EUROPEAN LANGUAGE EQUALITY

D2.13

Technology Deep Dive – Machine Translation

Authors	Aivars Bērziņš, Mārcis Pinnis, Inguna Skadiņa, Andrejs Vasiļjevs, Nora Aranberri, Joachim Van den Bogaert, Sally O'Connor, Mercedes García–Martínez, Iakes Goenaga, Jan Hajič, Manuel Herranz, Christian Lieske, Martin Popel, Maja Popović, Sheila Castilho, Federico Gaspari, Rudolf Rosa, Riccardo Superbo, Andy Way
Dissemination level	Public
Date	28-02-2022

About this document

Project	European Language Equality (ELE)
Grant agreement no.	LC-01641480 – 101018166 ELE
Coordinator	Prof. Dr. Andy Way (DCU)
Co-coordinator	Prof. Dr. Georg Rehm (DFKI)
Start date, duration	01-01-2021, 18 months
Deliverable number	D2.13
Deliverable title	Technology Deep Dive – Machine Translation
Type	Report
Number of pages	71
Status and version	Final
Dissemination level	Public
Date of delivery	Contractual: 28-02-2022 – Actual: 28-02-2022
Work package	WP2: European Language Equality – The Future Situation in 2030
Task	Task 2.3 Science – Technology – Society: Language Technology in 2030
Authors	Aivars Bērziņš, Mārcis Pinnis, Inguna Skadiņa, Andrejs Vasiļjevs, Nora Aranberri, Joachim Van den Bogaert, Sally O'Connor, Mercedes García–Martínez, Iakes Goenaga, Jan Hajič, Manuel Herranz, Christian Lieske, Martin Popel, Maja Popović, Sheila Castilho, Federico Gaspari, Rudolf Rosa, Riccardo Superbo, Andy Way
Reviewers	Gorka Labaka, Andy Way
EC project officers	Susan Fraser, Miklos Druskoczi
Contact	European Language Equality (ELE) ADAPT Centre, Dublin City University Glasnevin, Dublin 9, Ireland Prof. Dr. Andy Way – andy.way@adaptcentre.ie European Language Equality (ELE) DFKI GmbH Alt-Moabit 91c, 10559 Berlin, Germany Prof. Dr. Georg Rehm – georg.rehm@dfki.de http://www.european-language-equality.eu

Consortium

1	Dublin City University (Coordinator)	DCU	IE
2	Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (Co-coordinator)	DFKI	DE
3	Univerzita Karlova (Charles University)	CUNI	CZ
4	Athina-Erevnitiko Kentro Kainotomias Stis Technologies Tis Pliroforias, Ton Epikoinonion Kai Tis Gnosis	ILSP	GR
5	Universidad Del Pais Vasco/ Euskal Herriko Unibertsitatea (University of the Basque Country)	UPV/EHU	ES
6	CROSSLANG NV	CRSLNG	BE
7	European Federation of National Institutes for Language	EFNIL	LU
8	Réseau européen pour l'égalité des langues (European Language Equality Network)	ELEN	FR
9	European Civil Society Platform for Multilingualism	ECSPM	DK
10	CLARIN ERIC – Common Language Resources and Technology Infrastructure as a European Research Infrastructure Consortium	CLARIN	NL
11	Universiteit Leiden (University of Leiden)	ULEI	NL
12	Eurescom (European Institute for Research and Strategic Studies in Telecommunications GmbH)	ERSCM	DE
13	Stichting LIBER (Association of European Research Libraries)	LIBER	NL
14	Wikimedia Deutschland (Gesellschaft zur Förderung freien Wissens e. V.)	WMD	DE
15	Tilde SIA	TILDE	LV
16	Evaluations and Language Resources Distribution Agency	ELDA	FR
17	Expert System Iberia SL	EXPSYS	ES
18	HENSOLDT Analytics GmbH	HENS	AT
19	Xcelerator Machine Translations Ltd. (KantanMT)	KNTN	IE
20	PANGEANIC-B. I. Europa SLU	PAN	ES
21	Semantic Web Company GmbH	SWC	AT
22	SIRMA AI EAD (Ontotext)	ONTO	BG
23	SAP SE	SAP	DE
24	Universität Wien (University of Vienna)	UVIE	AT
25	Universiteit Antwerpen (University of Antwerp)	UANTW	BE
26	Institute for Bulgarian Language "Prof. Lyubomir Andreychin"	IBL	BG
27	Sveučilište u Zagrebu Filozofski fakultet (Univ. of Zagreb, Faculty of Hum. and Social Sciences)	FFZG	HR
28	Københavns Universitet (University of Copenhagen)	UCPH	DK
29	Tartu Ulikool (University of Tartu)	UTART	EE
30	Helsingin Yliopisto (University of Helsinki)	UHEL	FI
31	Centre National de la Recherche Scientifique	CNRS	FR
32	Nyelvtudományi Kutatóközpont (Research Institute for Linguistics)	NYTK	HU
33	Stofnun Árna Magnússonar í íslenskum fræðum SAM (Árni Magnússon Inst. for Icelandic Studies)	SAM	IS
34	Fondazione Bruno Kessler	FBK	IT
35	Latvijas Universitātes Matemātikas un Informātikas institūts (Institute of Mathematics and Computer Science, University of Latvia)	IMCS	LV
36	Lietuvių Kalbos Institutas (Institute of the Lithuanian Language)	LKI	LT
37	Luxembourg Institute of Science and Technology	LIST	LU
38	Università ta Malta (University of Malta)	UM	MT
39	Stichting Instituut voor de Nederlandse Taal (Dutch Language Institute)	INT	NL
40	Språkrådet (Language Council of Norway)	LCNOR	NO
41	Instytut Podstaw Informatyki Polskiej Akademii Nauk (Polish Academy of Sciences)	IPIPAN	PL
42	Universidade de Lisboa, Faculdade de Ciências (University of Lisbon, Faculty of Science)	FCULisbon	PT
43	Institutul de Cercetări Pentru Inteligență Artificială (Romanian Academy)	ICIA	RO
44	University of Cyprus, French and European Studies	UCY	CY
45	Jazykovedný ústav Ľudovíta Štúra Slovenskej akadémie vied (Slovak Academy of Sciences)	JULS	SK
46	Institut Jožef Stefan (Jozef Stefan Institute)	JSI	SI
47	Centro Nacional de Supercomputación (Barcelona Supercomputing Center)	BSC	ES
48	Kungliga Tekniska högskolan (Royal Institute of Technology)	KTH	SE
49	Universität Zürich (University of Zurich)	UZH	CH
50	University of Sheffield	USFD	UK
51	Universidad de Vigo (University of Vigo)	UVIGO	ES
52	Bangor University	BNGR	UK

Contents

List of Figures

List of Tables

List of Acronyms

AAN	Average Attention Transformer
AI	Artificial Intelligence
APE	Automatic Post-Editing
API	Application Programming Interface
ASICs	Application Specific Integrated Circuits
ASR	Automatic Speech Recognition
BLEU	BiLingual Evaluation Understudy
CAE	Correspondence Autoencoder
CAT	Computer-Assisted Translation
CEF	Connecting Europe Facility
CLM	Causal Language Modeling
CV	Computer Vision
CMS	Content Management System
DGT	Directorate-General for Translation
DIY	Do-It-Yourself
DL	Deep Learning
DLE	Digital Language Equality
DNN	Deep Neural Networks
DWA	Dynamic Weight Average
ELE	European Language Equality *(this project)*
ELE Programme	European Language Equality Programme *(the long-term, large-scale funding programme specified by the ELE project)*
ELG	European Language Grid (EU project, 2019-2022)
ELRC	European Language Resource Coordination
EU	European Union
FPGA	Field Programmable Gate Array
GDPR	General Data Protection Regulation
GPU	Graphics Processing Unit
HPC	High-Performance Computing
IPR	Intellectual Property Rights
LDA	Latent Dirichlet Allocation
LT	Language Technology/Technologies
LSP	Language Service Providers
META	Multilingual Europe Technology Alliance
META-NET	EU Network of Excellence to foster META
MIT	Massachusetts Institute of Technology
ML	Machine Learning
MLM	Masked Language Modeling
MMT	Multimodal Machine Translation

MT	Machine Translation
MTL	Multi-Task Learning
NER	Named Entity Recognition
NLP	Natural Language Processing
NMT	Neural Machine Translation
OCR	Optical Character Recognition
POS	Part-Of-Speech
qRAM	Quantum Random Access Memory
RRN	Recurrent Neural Network
SEO	Search Engine Optimisation
SME	Small and Medium-sized Enterprise
SMT	Statistical Machine Translation
ST	Speech Translation
TER	Translation Error Rate
UBDI	Unsupervised (Bilingual) Dictionary Induction
US	United States of America
WMT	Workshop of Machine Translation
WNGT	Workshop on Neural Generation and Translation
WP	Work Package

Abstract

Born from the dream that people can communicate freely with the help of computers, Machine Translation (MT) is one of the oldest language technologies being researched for more than 70 years. However, only during the last decade it has been widely accepted by the general public and in many cases it has become an indispensable tool for the global community, supporting communication between nations and lowering language barriers. This deliverable provides an overview of the current state of the art in the field of MT, offers technical and scientific forecasting for 2030, and provides recommendations for the advancement of MT as a critical technology for reaching the goal of European language equality.

Recently neural network models, in particular transformer models, opened up the possibility to build MT systems for many language pairs. However, quality and availability of MT solutions still greatly differs from language to language, as well as from domain to domain. There is still a major gap in technologies that can be successfully applied in under-resourced settings, can understand context and use common knowledge. Moreover, while text translation is widely used and available for many language pairs (but not for all domains), speech, multi-modal and sign language translation are still underdeveloped, are not widely used, and are available only for a rather small number of languages.

Unfortunately, today there is no common EU policy addressing language barriers. The absence of such roadmap and support for language technology (LT) at European level has resulted in a fragmented European market with uneven language support for the language communities of Europe. Aiming to achieve Digital Language Equality (DLE) in Europe by 2030, the ELE project is seeking to rectify this regrettable situation, including – but not exclusively – with regard to European languages being adequately served by MT. Insufficient investments in MT by European public and private sectors and disruptive power of global players have led to a weak and fragmented European MT industry. Dominance of global monopolies that often cross-subsidize MT services and provide them for free, increase European technological dependency and prevent the growth of European MT market. Usage of these solutions also includes significant risks related to protection of the data privacy and confidentiality. Taking into account the critical role that MT increasingly plays in the cross-language communication, Europe cannot afford the risk to depend on the global providers outside Europe.

The lack of usable and appropriate public language data in several European languages for MT engine training is a critical factor for lowering language barriers and enabling a truly multilingual Europe, where languages can thrive in the digital age. Publicly available multilingual data should include a greater diversity of domains and languages so that building high-quality MT systems becomes an option for all. There is also a disparity between publicly available and proprietary multi/bi-lingual corpora. A crucial breakthrough could be achieved if existing policy frameworks were adapted to make it mandatory for Member States to make all unprotected data in natural language-related workflows publicly available.

Training neural MT engines is resource intensive, requires massive infrastructure and has a heavy carbon footprint. By developing efficient models and hardware, the EU has the opportunity to be a pioneer in training and developing green LTs.

When looking ahead to 2030, we expect a major break-through towards efficient, omnipresent, high quality real-time translation between any European language pair and in any domain, regardless of the modality (written, spoken, sign language) of the input.

1. Introduction

Machine translation has been among the first application areas of natural language processing. Starting from the first attempts to apply dictionary-based approaches (Hutchins, 2004) till modern neural network systems, MT has aimed to provide automatic translation from one natural language into another.

Today, MT has become an important asset for multilingual Europe, allowing citizens, governments and businesses to communicate in their native language, breaking down language barriers and supporting the implementation of the European digital single market. For example, the eTranslation automated translation tool,[1] developed by the European Commission, and its various adoptions (e. g., EU Council Presidency Translator (Pinnis et al., 2021)[2]) provides reasonably good MT service in 24 EU official languages for governments, the public sector and SMEs.[3] However, in general, MT support and quality still differ from language to language, and from domain to domain. In particular, MT quality drops significantly when translation concerns less resourced languages, speech or terminology rich domains with limited parallel corpora and terminological resources available.

Limited MT support for less resourced language pairs and specialised domains from the European industry is among the reasons why European citizens still in many cases rely on global companies and external providers outside Europe (Vasiļjevs et al., 2019). The use of MT solutions developed by global companies outside Europe involves the risk that these freely available MT solutions could be taken away. Another question is on privacy of the data used and how this data is transmitted.

In 2012, the Language White Paper series (Rehm and Uszkoreit, 2012) presented a thorough analysis of LT support for 31 European languages. According to the White Papers for MT, in 2012 *Good support* only applied to English and *Moderate support* to only two widely spoken languages, namely French and Spanish, leaving the remaining 28 European languages of this study in clusters of *Fragmented* or *Weak or no support*.

Six years later, the lack of necessary LT support, including MT, and following significant disadvantage of less used languages was recognised by the European Parliament in the resolution on Language Equality in the Digital Age (European Parliament, 2018). In this resolution, the European Parliament "calls on administrations at all levels to improve access to online services and information in different languages, especially for services in cross-border regions and culture-related issues, and to use existing free and open-source LT, including MT, speech recognition and text-to-speech and intelligent linguistic systems, such as those performing multilingual information retrieval, summarising/abstracting and speech understanding, in order to improve the accessibility of those services".[4]

A recent review of LT support by ELE shows notable progress during the last decade for MT: while *Good support* still applies for only two languages, i. e., English and German, *Moderate support* is achieved for twelve languages, namely Catalan, Czech, Danish, Dutch, Finnish, French, Italian, Norwegian, Polish, Portuguese, Romanian, Spanish and Swedish.

Today translation technologies are widely used in various settings, including by language service providers (LSP), localization companies and businesses in Europe and across the world where content in client language is very important (e.g., hotels, travel, etc.). However, progress in MT is still often measured only in resource rich settings. Although there is on-going research on multilingual models, in practice there is a gap in technologies that can be successfully applied in under-resourced settings, to understand context and use common

[1] https://webgate.ec.europa.eu/etranslation/public/welcome.html

[2] e. g., EU Council Presidency Translator German Presidency https://www.eu2020.de/eu2020-en/presidency/uebersetzungstool/2361002

[3] As of February 2022, eTranslation is used by 108 projects – 87 projects reusing eTranslation and 21 project committed to analyse or reuse eTranslation.

[4] Article 44

knowledge. Moreover, even in resource rich settings, MT systems mostly perform sentence to sentence translation, ignoring context and the overall goal of specific communicative situations. Another challenge is various biases exposed in MT results, including gender, race, ethnicity and others.

While text-to-text translation is widely used today, speech, sign language and multi-modal MT is still relatively in its early stages. In fact, speech translation and voice interaction with devices are the key techniques to break the language barrier for human communication. Moreover, there is a growing need for translation of audiovisual content.

This deep dive report focuses on the MT landscape a decade after the publication of the Language White Papers, and offers an outlook into the future, up to 2030. The eight sections of this document analyse MT and its role in DLE, as well as its contribution to deep natural language understanding. The report analyses MT technology from several perspectives – main components and the current state of the art, main gaps, contribution to DLE and impact on society. It also outlines visions, breakthroughs needed and development goals for 2030.

Section 2 outlines the scope and key dimensions of this study (technologies and applications, modalities, data and language coverage). Section 3 provides insights into the main components of MT systems, describing neural MT architecture, in particular the Transformer model, as currently most popular and widely used MT architectures. In addition, it also discusses training data and presents various advanced techniques.

In Section 4 the current state of the art is discussed. The section starts with a short overview of MT history, followed by an introduction to the current MT research approaches and benchmarks that enable linguistic equality through multilingual MT models. The most important industrial achievements and several neural MT solutions provided by companies are reviewed. Five important application areas (i. e., online MT for general use, video games, health, public administration and legal documentation, and e-commerce) that have a high impact on society are analysed in more detail.

Section 5 identifies the main gaps in four important dimensions of MT, namely data, technology, approaches and legal aspects (regulation). Besides problems of data availability and usability and need for less-resourced technologies, gaps and current limitations related to multi-modal MT are highlighted. Moreover, several challenges in localization workflows are identified. Finally, gaps related to the legislation are discussed, covering not only copyright and licensing issues, but also lack of common policy addressing language barriers, dependence from large global companies and heavy carbon footprint coming from AI and MT research and development activities.

Section 6 discusses the contribution of MT to DLE and its impact on society (including privacy, GDPR and IPR). It stresses importance of MT in knowledge acquisition or information access to low resourced languages, while emphasising MT's contribution to DLE in general, and especially for the languages of Europe. This section also discusses several everyday use cases that have an impact on society and illustrate MT use in different spheres.

Section 7 lists needed breakthroughs in MT related to system development (including interoperability, explainability, contextualisation, hardware needs and opportunities and offers from quantum computing), data collection and EU policies, focusing on carbon-neutral and trustworthy AI. Special attention is paid to the GDPR regulation and its limitations and impact on MT and the LT industry, as well as research and development activities.

Section 8 provides insights into the main technology visions and development goals for MT from four perspectives – systems and models, data, evaluation and applications. Besides addressing gaps identified in Section 5, this section also discusses the need for ethical and fair MT, stresses the importance of smaller MT models and expanding MT to other natural language processing (NLP) tasks and application areas. Finally, the process, current strategies and methods for evaluation, as well as the needs and visions for future evaluation metrics are discussed, given their significant strategic importance, both for the MT community as a whole and, ultimately, for the end-users.

Finally, Section 9 discusses the contribution of MT to deep natural language understanding by stressing the need for LTs that go beyond simple phrase or sentence understanding, are aware of context and thus allow MT systems to generate output which is faithful to the intended communication, regardless of its complexity and non-verbal or implicit components.

2. Scope of this Deep Dive

MT is one of the NLP fields that has received attention for many years. It has been analysed, criticised and praised from different perspectives and in different contexts. The scope of this deliverable is to analyse progress in MT (current state of technologies), identify the main gaps and outline visions and development goals in this field towards DLE and deep natural language understanding by 2030. Following the common methodology of all WP2 deep dives, a multidimensional approach was applied, consolidating views of the European research and industrial stakeholders. Different modalities of MT, including text, speech and audiovisual, are covered. Within the scope of this deliverable, we look closer at MT technologies and approaches, their applications, data necessary for training and evaluation, infrastructures required, legal and ethical aspects. These topics are analysed from the perspective of DLE and in the context of European needs and aspirations for the future.

The modern **neural network MT techniques** are considered as the current state of the art in this this deliverable. The most important and promising research approaches (e.g., multilingual MT, unsupervised MT) and benchmarks are analysed not only from the readiness and quality aspects, but also in terms of language coverage and their contribution to DLE.

We look at **the current services and technologies** offered by MT providers in the European market. This includes free online translation services from global technology companies and smaller niche players, commercial online MT services (usually based on a subscription model) and customised MT solutions. The dominance of global companies in the free online translation market and risks for Europeans caused by this dependence are among the key topics discussed in this deliverable, especially to identify solutions going forward.

In addition to general-purpose MT, we also consider **specific domains and applications** where MT technologies are particularly important, such as public administration, governmental and EU services, the translation and localization industry, e-Commerce, computer-mediated communication, education, tourism, etc.

We also include recent **research breakthroughs** associated with MT. In this regard, special emphasis is placed on deep learning architectures (e.g., Transformer models), research on NMT model efficiency, use of broader contexts (e.g., documents instead of isolated sentences) and multiple source inputs (e.g., source sentences in multiple languages), use of linguistic knowledge (e.g., morphology, syntax, semantics) and external knowledge (e.g., domain-specific terminology, domain information, etc.), multi-lingual and multi-domain NMT, use of pre-trained models (e.g., BERT, mBART, etc.), multi-task learning, automatic post-editing, and other methods that allow achieving state-of-the-art translation quality for NMT systems.

Another important area of research that we cover concerns leveraging other **modalities** of information in addition to text, such as speech and visual modalities. We discuss the latest achievements in speech translation and multi-modal MT.

We also discuss MT research topics that address **specific needs of various MT use cases** in translation and localisation services, such as interactive MT, automatic quality estimation, and online learning.

Apart from being able to develop MT systems, it is as important to be able to show whether the systems are state-of-the-art and whether they are fit for purpose. Therefore, the document also discusses MT evaluation metrics (both automatic and manual metrics), as well as metrics that are used in use cases to show the benefits of using MT.

While MT technologies today are available for most of the European languages, many of these languages are less attractive from a business point of view, and consequently they are not so well equipped with MT tools and other widely used NLP techniques. Throughout this document, **language coverage** is addressed as another key dimension for DLE.

Another important aspect towards language equality is the **availability of data** necessary for MT training and methods allowing to overcome data scarcity for less and low resource languages and domains. Although different machine learning techniques try to reduce the overall dependence on data, data-sets are still indispensable for MT training and data deserves special attention in the context of **less resourced languages and domains**. We discuss the latest research achievements of unsupervised and low-resource MT.

This deliverable also discusses **legal and ethical aspects** related to the development, production and use of MT systems and services. It analyses IPR (and GDPR) restrictions and the Fair use principle from the developer's perspective, and privacy and security issues from the user's perspective.

Finally, all the above-mentioned aspects are taken into consideration from the perspective of their impact on society, with a focus on Europe. The document also provides a series of **recommendations** on how to address the current limitations of MT technologies and on how to contribute to DLE as a crucial goal for Europe and its citizens.

3. Machine Translation: Main Components

Different MT types have been described (e. g., rule-based, example-based, statistical phrase-based, hierarchical), but here we will focus on the recent development of Neural MT (NMT) only, citing from an overview of Popel (2018). We describe the main MT components: general NMT architecture (Section 3.1) and the currently most popular example – Transformer (Section 3.2), training data (Section 3.3), and various advanced techniques: backtranslation (Section 3.4) and checkpoint averaging (Section 3.5).

Of course, there are also many other components related to MT, which are not described in this deliverable, for example: automatic speech recognition (ASR, see, e. g., reports of the ELITR project[5]) and text-to-speech (TTS), which is needed in the speech-to-speech translation pipeline; cross-lingual information retrieval; multilingual summarization (see again, e. g., the ELITR project); integration into production systems and multilingual websites using suitable metadata formats.[6]

3.1. General NMT Architecture

3.1.1. Early NMT approaches

There were several early attempts to exploit *neural networks* (at that time also known as *connectionist models* or *continuous-space models*, in some contexts) in MT (e. g. Chrisman, 1991; Waibel et al., 1991; Forcada and Ñeco, 1997; Castaño et al., 1997). However, the era of (modern) end-to-end NMT systems (Kalchbrenner and Blunsom, 2013; Sutskever et al., 2014; Bahdanau et al., 2014a) started about two decades later, when the computational resources (GPU) became capable of training large models. NMT systems became competitive in well-known shared tasks: WMT 2015 (Jean et al., 2015), IWSLT 2015 (Luong and Manning, 2015) and WMT 2016 (Sennrich et al., 2016c). Most of these systems use recurrent (RNN) layers (e. g., LSTM (Hochreiter and Schmidhuber, 1997) or GRU (Cho et al., 2014)), though some use convolutional layers as well (Kalchbrenner and Blunsom, 2013; Gehring et al., 2017). Vaswani

[5] https://elitr.eu
[6] https://www.w3.org/TR/mlw-metadata-us-impl

et al. (2017) introduced a novel model called Transformer, which uses *self-attention* instead of the recurrent or convolutional layers.

Transformer outperformed all the above-mentioned models (Vaswani et al., 2017) and almost all systems in recent WMT shared tasks are based on Transformer.[7] We thus describe the model in more detail (Section 3.2), after summarising the basic principles of NMT architectures (Section 3.1). Finally, we explain what we mean by the latent structures emerging in Transformer self-attention layers (Section 3.2.5).

3.1.2. Token representation

In NMT, each input sentence is first tokenised into a sequence of tokens. The early NMT systems used words as tokens (Sutskever et al., 2014), which resulted in large vocabularies and a necessity to handle unknown words, referred to as out-of-vocabulary items. An alternative is to use characters as the tokens (Lee et al., 2016) or to use a hybrid word-character approach (Luong and Manning, 2016). However, the most popular approach today is to split words into subword units (*subwords*)[8] and use these as the tokens for NMT (Sennrich et al., 2016b). The subword vocabulary is trained so that frequent words are represented with a single subword, while rarer words are encoded into multiple subwords. There are several algorithms for training subword vocabularies (Schuster and Nakajima, 2012; Sennrich et al., 2016b; Macháček et al., 2018; Kudo, 2018).

Each token is represented as a real-value vector, called *embedding* (word embedding, subword embedding or character embedding). These embeddings can be pre-trained on monolingual texts (e. g., with `word2vec` (Mikolov et al., 2013)), but most NMT systems initialise them randomly and train them jointly with the whole translation.

3.1.3. Encoder-decoder architecture

Most NMT systems are based on an *encoder-decoder* architecture. The encoder maps the input sequence to a vector of *hidden states* (sometimes called *continuous representation* or *sentence embedding*). The decoder maps the hidden states into the output sequence (of target-language tokens). Each hidden state usually corresponds to one position (token) in the input sequence, so in general, the vector of hidden states has a variable length.[9] The early NMT systems (Sutskever et al., 2014) used only the last hidden vector as an input for the decoder. Thus, the training was forced to encode all the information about the input sentence into a fixed-length vector. Bahdanau et al. (2014a) suggested using a bidirectional GRU (gated recurrent unit) in the encoder. More importantly, they introduced an *encoder-decoder attention* mechanism, where the decoder has access to all of the encoder's hidden states. This way, when generating each output token, the decoder can *attend* to different parts of the input sentence. The encoder-decoder attention mechanism circumvents the fixed-length sentence-representation restriction and improves the translation quality, especially on longer sentences (Bahdanau et al., 2014a).

[7] See, e. g., appendix C in https://www.statmt.org/wmt21/pdf/2021.wmt-1.1.pdf.

[8] For example, the German word *Forschungsinstituten* (i. e., research institutes) is encoded with three subwords: *Forsch* + *ungsinstitu* + *ten_*.

[9] For practical reasons of (mini-)batch training on GPU, sentences within one batch are usually padded to a fixed length according to the longest sentence in the batch.

3.1.4. Inference

The process of translating sentences (at test time) with a trained NMT model is usually called *inference*.[10] Most NMT systems use *auto-regressive* inference.[11] This means that the output sentence is generated token by token and after each token is generated, its embedding is used as an input for generating the next token. In case of an RNN decoder, the decoding time grows linearly with the sentence length. The decoding finishes once the decoder generates a special end-of-sentence token.

3.1.5. Training

The advantage of NMT systems is that all their components can be trained in an end-to-end fashion. This is in contrast with SMT and TectoMT, where most components had to be trained separately. NMT is usually trained using backpropagation optimising the *cross-entropy* loss of the last decoder's *softmax* layer, which predicts output token probabilities,[12] but there are also NMT systems optimising sentence-level metrics (e. g., BLEU or simulated human feedback) with reinforcement learning techniques (e. g. Nguyen et al., 2017).

Importantly, NMT usually uses the *teacher-forcing* technique: when generating the next word during training, i. e., it uses the previous word from the reference translation as the input instead of using the previously predicted word.

3.2. Transformer

Transformer (Vaswani et al., 2017) follows the general encoder-decoder architecture as described above, but instead of RNN layers it uses *self-attention* and feed-forward layers. This allows to speed up the training and partially also the decoding thanks to better usage of parallelism.[13]

3.2.1. Attention

In general terms, *attention* can be defined as a function mapping three vectors of queries (Q), keys (K) and values (V) to an output vector, which is a weighted sum of the values V. The weight is computed as a compatibility of the corresponding key and query. See Appendix A.1 for details.

In the *encoder-decoder attention*, keys and values come from the encoder's topmost layer and queries come from the decoder's previous layer.

In *self-attention*, all queries, keys and values come from the output of the previous layer. Self-attention is used both in encoder and decoder, but in decoder it is *masked*, so each position attends only to preceding positions, because the following positions will not be known at inference time due to the autoregressive property of the decoder.

[10] In statistical MT (SMT), it is usually called *decoding*, but this is a slightly ambiguous term in NMT, because it can mean either the inference or only using the decoder (possibly at training time).

[11] There are also experimental systems with non-autoregressive inference (e. g. Gu et al., 2017; Zhang et al., 2018a; Roy et al., 2018). Their inference is faster, but the translation quality has not achieved the level of autoregressive models yet.

[12] $softmax(\mathbf{x})_j = exp(x_j)/\sum_i exp(x_i)$ and the *cross-entropy* loss is $H(\mathbf{y}, \mathbf{p}) = \sum_i y_i log(p_i)$, where $\mathbf{p} = softmax(\mathbf{x})$ is the predicted probability distribution of output tokens and $\mathbf{y}$ is the training-data "distribution", which is usually a *one-hot* vector ($y_k = 1$ for the token k on a given position in the reference translation, all other tokens $i \neq k$ have $y_i = 0$).

[13] During training both encoder and decoder work in a non-autoregressive mode, that is all positions are encoded/decoded in parallel. During inference, only the encoder works non-autoregressively.

It is possible to use a single self-attention function in each layer, but the translation quality is improved (Vaswani et al., 2017) when combining multiple *attention heads*, so that each head can focus on different phenomenon.

3.2.2. Encoder and decoder

The encoder of Transformer consists of 6 stacked layers of identical form. Each layer has two sublayers (multi-head self-attention and position-wise feed-forward sublayer), a residual connection around each of the sublayers and a layer normalization. See Appendix A.3 for details.

The decoder is similar to the encoder, but in addition to the self-attention and feed-forward sublayers, each of the 6 layers includes also the encoder-decoder attention sublayer.

3.2.3. Positional encoding

Transformer contains no recurrence nor convolution and thus the information about position of the tokens in the sentence must be supplied by other means. Transformer encodes the absolute position in the sequence into *positional encoding* vector, which is subsequently summed with subword embeddings and provided as the input to the first layer of the encoder. See Appendix A.4 for details.

An alternative solution is to extend the self-attention formula with a term which depends on the relative distance of the key and query (Shaw et al., 2018).

3.2.4. Further reading

For more details on Transformer, see the original paper (Vaswani et al., 2017). There is also a blog explaining Transformer with many illustrations, showing, e. g., a visualization of the positional encoding.[14] Finally, Popel and Bojar (2018) provide more information about training Transformer models.

3.2.5. Latent Structures in Transformer

The attention (Equation 1) is based on a "compatibility" function $softmax\left(QK^T d_k^{-0.5}\right)$, which assigns a weight $w_{i,j} \in \langle 0, 1\rangle$ to each query-key pair (Q_i, K_j). In the case of the multi-head self-attention in the encoder, the queries and keys are different projections of the vectors representing each token on a given layer. For a given sentence, encoder's layer and head, it is thus possible to visualise this self-attention as a weighted bipartite graph, where the weight of each edge is defined by $w_{i,j}$.

Figure 1 shows an example of such visualisation for different heads. Each head is visualised in a different colour and edge weight is indicated by thickness. Each of the three subfigures shows another attention head in encoder layer 5 (out of 6). The words in the left column in each of the three visualisations represent vectors corresponding to these words on the input to the fifth layer of the encoder. By copying these vectors multiplied by the $w_{i,j}$ weights, Transformer starts building the input for the sixth (and last) layer of the encoder. Thus we can imagine that the words in the left column represent the sixth layer.[15] The rightmost subfigure shows two attention heads, but focusing only on the word '*its*' and illustrating coreference resolution.[16]

[14] https://jalammar.github.io/illustrated-transformer/

[15] As described in Section 3.2, self-attention is followed by the PFFN sublayer and both sublayers are wrapped by residual skip connections and layer normalization. However, self-attention is the only component, which combines representations on different positions.

[16] Incidentally, all words in the example sentence are frequent enough to be encoded with a single subword.

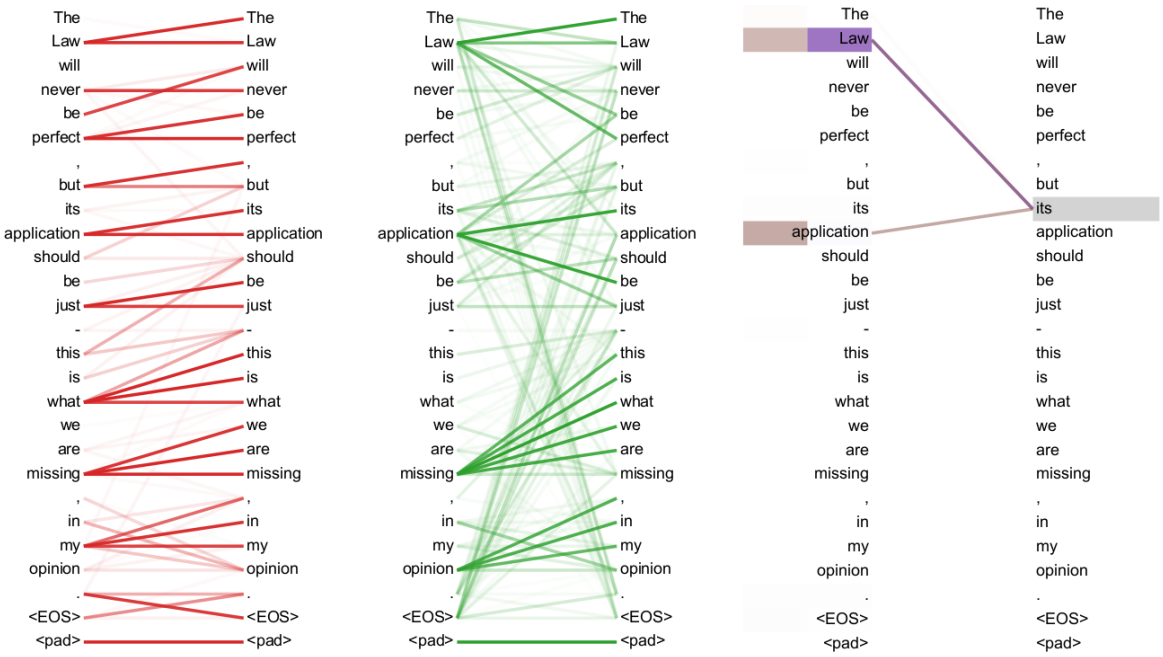

Figure 1: Visualization of self-attention in a Transformer model trained on English→German translation. Adapted from Vaswani et al. (2017).

We can see several interesting phenomena in Figure 1:

- The self-attention has a relatively sharp distribution – each position (in the right column) attends to a small number of positions (in the left column). Often, most of the attention focuses on a single position.
- Each head obviously specialises in a different task. The red-marked head in Figure 1 focuses mostly on short-distance dependencies and short phrases. The green-marked head focuses on longer-distance dependencies and longer phrases. The violet-marked head resolves the *its–Law* coreference link. Intuitively, this makes sense – by copying the vector representation of "*Law*" to the position of "*its*", we allow the further layers (in encoder and decoder) to "understand" the meaning of "*its application should be just*" and translate it correctly within a given context.
- The visualised self-attention structures resemble syntactic and semantic structures that are present in manually annotated treebanks. However, the self-attention structures emerged in the end-to-end training of English→German translations, where no linguistic annotations were provided in the training data. We call these structures *latent* because they correspond to latent variables of the Transformer model, i. e., variables which are not (explicitly) visible in the data, but need to be inferred.

3.3. Training data

The quality of NMT depends heavily on the amount and quality of the training parallel sentences. Although millions of parallel sentences are freely available for several language pairs,[17] the available material is still not enough for achieving optimal results in many Eu-

[17] http://opus.nlpl.eu provides 32 English-X parallel corpora larger than 500M tokens for the following 17 languages: ar, bg, cs, de, el, es, fr, hr, hu, it, nl, pl, pt, ro, ru, sr, tr.

ropean languages. Moreover, the parallel data may not be available for the target domain (e. g., medical or IT domain). A common solution is to use monolingual target-language data, which is usually available in much larger amounts than the parallel data. The current best practice in improving the quality of NMT using monolingual target-language data is *back-translation* (Sennrich et al., 2016a), where the monolingual data is machine-translated to the source language, and the resulting sentence pairs are used as additional (*synthetic*) parallel training data as described in the following section.

3.4. Backtranslation

In SMT, target-language monolingual data are typically used only to build language models, but there are also works showing that additional synthetic parallel training data created with backtranslation can improve SMT (Schwenk, 2008; Bertoldi and Federico, 2009; Bojar and Tamchyna, 2011).

The early end-to-end NMT systems (e. g., Bahdanau et al. (2014a); see Section 3.1) used no target-language monolingual data. The first attempts to exploit this abundant data to improve NMT used a separately-trained RNN language model (Gülçehre et al., 2015) or sentence pairs with dummy (empty) source side used to train only the NMT decoder (Sennrich et al., 2016a). However, synthetic parallel data created by backtranslation quickly became popular because this approach is very easy to use with any NMT architecture and it gives better results than both aforementioned approaches (Sennrich et al., 2016a). So, while in SMT, target-language monolingual data are typically used for building language models, in NMT they are typically used for building synthetic parallel data via backtranslation.

Sennrich et al. (2017) compared two regimes of how to incorporate the synthetic training data. In the *fine-tuned* regime, a system is trained first on the authentic parallel data and then after several epochs it is trained on a 1:1 mix of authentic and synthetic data. In the *mixed* regime, the 1:1 mixed data is used from the beginning of training. In both cases, the 1:1 mix means shuffling the data randomly at the sentence level, possibly oversampling the smaller of the two data sources. Popel et al. (2020) introduced a third regime: *block backtranslation*, where the authentic and synthetic parallel data are simply concatenated (without shuffling), which is synergically combined with checkpoint averaging (see Section 3.5).

Note that although the data is termed *synthetic*, it is only its source side which is machine-translated. The target side is authentic and thus can improve the fluency (and sometimes even adequacy) of the final translations, simply by increasing the total size and diversity of the training data.[18] Backtranslation can also be used as a domain-adaptation technique if the target-language monolingual data is in-domain or filtered to match the target domain (Moore and Lewis, 2010).

3.5. Checkpoint Averaging

A popular way of improving the translation quality in NMT is ensembling, where several independent models are trained and during inference (decoding) each target token is chosen according to an averaged probability distribution (using argmax in the case of greedy decoding) and used for further decisions in the autoregressive decoder of each model.

However, ensembling is expensive both in training and inference time. The training time can be decreased by using *checkpoint ensembles* (Sennrich et al., 2017), where N last checkpoints of a single training run are used instead of N independently trained models. Checkpoint ensembles are usually worse than independent ensembles (Sennrich et al., 2017), but

[18] Rarrick et al. (2011) show that it is beneficial to filter out machine-translated sentences from the training data, but this concerns target-side synthetic sentences, so it does not contradict the improvements brought by back-translation.

allow to use more models in the ensemble thanks to shorter training time. The inference time can be decreased by using *checkpoint averaging*, where the weights in the N last checkpoints are element-wise averaged, creating a single averaged model.

Averaging weights of independently trained models (with different random initialization, including embedding weights) does not work because the model weights are not "compatible". Utans (1996) suggests to train an initial network on all data and then use it as a starting point for training N networks on different subsets of the data and subsequently average weights of these networks. So similarly to checkpoint averaging, the averaged models share the same initial training and are not completely independent.

Checkpoint averaging has been first used in NMT by Junczys-Dowmunt et al. (2016, § 6.3), who report that averaging four checkpoints is "not much worse than the actual ensemble" of the same four checkpoints and it is better than ensembles of two checkpoints. Averaging ten checkpoints "even slightly outperforms the real four-model ensemble".

Checkpoint averaging is popular in recent NMT systems (e. g. Vaswani et al., 2017) because it has almost no additional cost (averaging takes a few minutes), the results of averaged models have lower variance in BLEU and are almost always at least slightly better than without averaging. The interplay of checkpoint averaging with training dynamics is still not fully explored and understood. An example of recent surprisingly promising improvements is *stochastic weight averaging* (Izmailov et al., 2018), where checkpoint averaging is used in combination with *cyclic learning rate* (Loshchilov and Hutter, 2016) or constant learning rate, which interestingly leads to faster convergence and better generalisation, as it finds much broader optima than standard SGD (stochastic gradient descent).

4. Machine Translation: Current State of the Art

4.1. Current MT research approaches and benchmarks

The application of neural networks to MT has opened the path to developing a universal engine whose ultimate goal is to train a single model to translate between an arbitrary language pair. As a supervised system, to attain this goal, the architecture must model a massively multi-way input-output mapping task under strong constraints: a huge number of languages, different scripting systems, heavy data imbalance across languages and domains, and a practical limit on model capacity.

The effects of different advanced approaches for multilingual MT models have been investigated by Yang et al. (2021), for example. They first explore the way to leverage the pre-trained language models that have been trained with large-scale monolingual data using the publicly available DeltaLM-Large multilingual pre-trained encoder-decoder model (Ma et al., 2021) to initialise the model. For efficient training, they apply progressive learning (Li et al., 2020; Zhou et al., 2021; Zhang et al., 2020) to their model that continue-trains a shallow model into a deep model. Specifically, they first train a model with 24 encoder layers, and then continue-train it by adding 12 layers on top of the encoder. Additionally, they implement iterative back-translation (Hoang et al., 2018; Dou et al., 2020) that translates the data for multiple rounds for data augmentation. They evaluate the system on the publicly available Flores-101 data set (Goyal et al., 2021) consisting of 3001 sentences extracted from English Wikipedia, covering a variety of different topics and domains, and translated into 101 languages by professional translators through a carefully controlled process. While the results are very promising, they reflect a worrying trend: when English is involved in the translation process either as a source or target language, the BLEU scores are rather high (33.35 points on average when translating into English and 27.39 points on average out of English). However, the results worsen considerably when only translation in language pairs without English is considered (17.44 points on average).

Along similar lines, Chen et al. (2021) build a single multilingual translation system with a hypothesis that a universal cross-language representation leads to better multilingual translation performance. They explore three types of back-translation methods, i. e., beam search with the beam size of five (Sennrich et al., 2016a), unconstrained sampling (Edunov et al., 2018) and sampling constrained to the most 10 likely words (Graves, 2013; Fan et al., 2018). Besides, they also explore the effect of including vocabularies and varying volumes of synthetic data during training. Surprisingly, smaller vocabularies perform better, and the extensive monolingual English data only offers a modest improvement. Interestingly, they obtained 34.96 and 33.34 average sp-BLEU scores (Goyal et al., 2021) for five Central/East European languages and English (30 directions) and five South East Asian languages and English (30 directions), respectively. This shows that techniques which favour languages with a differing configuration to English are accessible and research could be directed to working on such languges.

A key aspect to providing multilingual translation is scaling and multiple attempts at testing scalability limits have been carried out. One of the best results were obtained by Tran et al. (2021b), who trained two multilingual systems for 14 language directions: English to and from Czech, German, Hausa, Icelandic, Japanese, Russian, and Chinese. In addition to well-known techniques such as large-scale back-translation (Edunov et al., 2018), in-domain fine-tuning (Tang et al., 2020) and noisy channel re-ranking (Yee et al., 2019), the researchers also experimented with scaling dense transformers (Vaswani et al., 2017) (up to 4.7B parameters) and sparse Mixture-of-Expert (Jacobs et al., 1991; Shazeer et al., 2017) (up to 52B parameters). Compared to previous year's winning submissions on the Workshop on Statistical Machine Translation, their multilingual system improved the translation quality on all language directions, with an average improvement of 2.0 BLEU. This type of research is proving that working with multiple languages at a time – rather than using the traditional bilingual approach – is beneficial for the final systems' quality.

Yet, researchers are also exploring alternative approaches to multilingual translation. One such strategy, which uses a new combination of existing techniques and has reported highly successful results is that proposed by Zeng et al. (2021). They develop a system that is based on the Transformer (Vaswani et al., 2017) with some effective variants, such as different combinations of Average Attention Transformer (AAN) models (Zhang et al., 2018b), weighted-aan model (Ahmed et al., 2017) and talking-heads attention model (Shazeer et al., 2020). In addition they employ data selection, Large-scale Back-Translation (Edunov et al., 2018) and Knowledge Distillation (Kim and Rush, 2016) for synthetic data generation. They also employ Target Denoising (Meng et al., 2020), Graduated Label-smoothing (Wang et al., 2020) and Confidence-Aware Scheduled Sampling (Liu et al., 2021) for advanced finetuning and a self-Bleu based model ensemble (Meng et al., 2020). Their systems achieve 36.9, 46.9, 27.8 and 31.3 case-sensitive BLEU scores on English-to-Chinese, English-to-Japanese, Japanese-to-English and English-to-German, respectively. Yet again, we see another example where English is involved in the translation process and high BLEU scores are obtained. It is worth mentioning, nonetheless, that pairing languages are, although full-resourced, largely different in terms of typology, which might shed light on the technical difficulties posed by matching those language families.

If we turn to the goal of achieving linguistic equality, which seems to be one of the main weaknesses of both research projects and industrial provision, one of the most interesting approaches is unsupervised MT (Artetxe et al., 2018; Lample et al., 2018) where no bilingual parallel data is needed to train a fully working system. Conneau and Lample (2019) showed for the first time the strong impact of cross-lingual language model (XLM) pretraining, investigating two unsupervised training objectives that require only monolingual corpora: Causal Language Modeling (CLM) (Radford et al., 2018) and Masked Language Modeling (MLM) (Devlin et al., 2018). They show that both approaches provide strong cross-lingual features that can be useful for pretraining models. On unsupervised MT, they show that MLM

pretraining is extremely effective reaching a new state of the art of 34.3 BLEU on WMT'16 German-English, outperforming the previous best approach by more than 9 BLEU points. They also demonstrate that cross-lingual language models can be used to improve the perplexity of low-resourced languages (Nepali language model). Without using a single parallel sentence, a cross-lingual language model fine-tuned on the XNLI cross-lingual classification benchmark (Conneau et al., 2018) already outperforms the previous supervised state of the art by 1.3% accuracy on average.

In recent years, researchers have pushed this approach so that it is slowly catching up with the translation quality obtained by supervised systems. For instance, Han et al. (2021) show state-of-the-art unsupervised neural MT system derived from a generatively pre-trained language model. Their method is a concatenation of three steps: few-shot amplification, distillation, and back-translation (Sennrich et al., 2016a). They first use the zero-shot translation ability of a large pre-trained language model (GPT-3) (Brown et al., 2020) to generate translations for a small set of unlabeled sentences. In the next step they amplify these zero-shot translations by using them as few-shot demonstrations for sampling a larger synthetic data set. This data set is distilled by discarding the few-shot demonstrations and then fine-tuning. During back-translation, they repeatedly generate translations for a set of inputs and then fine-tune a single language model on both directions of the translation task at once, ensuring cycle-consistency by swapping the roles of gold monotext and generated translations when fine-tuning. By using their method to leverage GPT-3's zero-shot translation capability, they obtain a new state-of-the-art in unsupervised translation on the WMT14 English-French benchmark, attaining a BLEU score of 42.1. While still restricted to a main-stream, full-resourced language pair, learning outcomes are promising for lower-resource pairs.

While developing high quality, robust systems is the main goal of MT research, the added value of Automatic Post-Editing (APE) should not be disregarded. APE involves automatically correcting MT outputs before they are addressed by human language reviewers. What is more, along with fixing systematic errors in the output, APE models have the ability to adapt general purpose MT systems to new domains, registers, and so on always with the final aim of obtaining better translations but also that of reducing the human post-editing effort (Chatterjee et al., 2015). APE has seen significant progress with transformer-based models (Yang et al., 2020; Lopes et al., 2019) dominating the landscape as opposed to the earlier statistical-based models (Simard et al., 2007; Bechara et al., 2012) and RNN based sequence-to-sequence models (Junczys-Dowmunt and Grundkiewicz, 2017).

Such an example is that by Oh et al. (2021) whose APE system is built based on Facebook FAIR's WMT19 News Translation Model (Ng et al., 2019) and is post-trained on WMT21 News-Translation Data (Koehn et al., 2005) and artificial synthetic data (Junczys-Dowmunt and Grundkiewicz, 2016) with Curriculum Training Strategy (Xu et al., 2020). For finetuning, Multi-Task Learning (MTL) (Ruder, 2017) is applied with related NLP subtasks such as Part-Of-Speech (POS), Named Entity Recognition (NER), Masked Language Model (MLM), and Keep/Translate are added to the model to reduce the over-fitting as well as achieve better performance. For better training efficiency, the Dynamic Weight Average (DWA) mechanism (Liu et al., 2019) is applied during the MTL to keep the correct balance between these subtasks. Their experimental results show that their model is able to effectively detect and correct the errors made by a high-quality NMT system, improving the score by -2.848 and +3.74 on the development dataset (English-German) in terms of TER and BLEU, respectively.

What is clear from a review of the latest research efforts is that the use of notably deep learning techniques has proven a major boost in the area and a decisive step forward in the quality MT systems provide. Multilingual systems, which are able to leverage knowledge and data from the different languages involved, are becoming the most popular frameworks. Nonetheless, the focus on big, fully-resourced languages, and English in particular, to test new techniques is concerning as it works against diversity, reinforcing already-existing disparities. Fortunately, a novel approach attracting the attention of many researchers is un-

supervised MT, where (every time less) monolingual data suffices to build a working system. While much work remains to be done in this area, together with universal MT, it emerges as one of the key pillars to drive language equality.

4.2. MT adoption and provision in the language industry

Today, most companies that base their business on providing translation solutions employ neural networks in their MT systems. A quick look at providers' solutions gives a clear overview of the strengths of each company, as well as the issues that remain relevant for the industry regarding the successful implementation of the technology.

A key aspect that most companies emphasise is the capacity for domain adaptation. This allows for specialised engines that learn from domain-specific texts, and therefore, are aware of specialist terminology and phraseology, avoiding the noise that words and expressions from other fields might introduce in the learning process. Here we find companies that work exclusively for certain areas (e. g., Lingua Custodia) as well as companies who develop separate systems by domain (e. g., Lengoo, Pangeanic, RWS, Welocalize).

Further customisation is also highly valued and widely advertised. This can take several forms. Most frequently, MT providers have their own generic or domain-specific engine which is retrained and refined with customer's own data. This often includes not only large amounts of parallel texts (previous translations), but also self-improving glossary technology (e. g., Across, Language I/O). Alternatively, do-it-yourself (DIY) MT opportunities are also provided where customers have full control over system choices and build their own from scratch using solely their company data (previous translations, terminology lists, etc.). Customers can then choose to host the MT systems internally in a private cloud or hire third-party resources for maintenance. What is also important is that such solutions can be made accessible in all kinds of devices, from computers to smart phones or tablets.

Seamless integration of MT within existing localisation workflows is paramount for the successful adoption of the technology. Companies go to great lengths to provide solutions, often in the form of plug-ins, that can be easily integrated in the client's computer-assisted translation (CAT) tool, content management system (CMS) or website to avoid changes in the main frameworks and allow for their usual communication channels with their customers (e. g., Kantan, Pangeanic, Tilde, Unbabel).

The text type involved in the translation process is also distinctive across companies. For example, we see companies specialising in email, article and chat translation, and therefore, pushing for real-time adaptive MT, that is, engines that work fast and adjust the output as conversations progress based on previous interactions (Language I/O), while others emphasise multimodality.

MT is coupled with post-editing when a level of accuracy and/or cultural adaptation is required in the final translated text. Some companies highlight their platform's functionalities to address this, be it directed at professional translators or crowd-sourcing platforms (e. g., Lengoo GmbH, Unbabel).

Interestingly, privacy and security emerge as matters of utmost interest in the industry. Text submitted for translation may include sensitive product or customer information and clients are often reluctant to hand these details over to third-party technology provides, make them available to external post-editors and even to the MT systems, which can learn from edits made to the raw output. The lack of understanding of how MT works and the unclear legal rights, obligations and consequences of misuse have clients seeking solutions backed with specific privacy and security functionalities (e. g., Across, Language Weaver, Pangeanic).

To conclude this list of features that companies highlight in their MT solutions, it is interesting to note several less representative yet interesting initiatives such as scalability concerns (e. g., KantanMT, Lilt), the use of open-source technology (e. g., Pangeanic, Apertium) and

speech MT (e. g., Papercup).

There are myriads of LSPs, each with their own strengths and limitations. However, it is undeniable that it is tech giants such as Amazon, Google and Microsoft, or big multinational LSPs such as Lionbridge and Welocalize who set the standards and best practices for LT development and provision. Most such companies are headquartered outside Europe and, consequently, have business and societal objectives that do not always align with European needs and goals. The dominance of those global companies exposes Europe's lack of market power which results in increasing market disparities. The absence of a clear roadmap and support for LT at European level translates into an incohesive, fragmented European market with disparate language support for the language communities of Europe.

4.3. Specialised MT: applications that change society

We stopped a long time ago seeing MT as a topic exclusive to research circles and even as a mere tool to decrease manual translation turnaround time for LSPs. Thanks to the advances of the last few years, MT is starting to play a key role in real communication activities across the globe. People are consuming multilingual content at an increasing pace, across many platforms and digital media, and both professional and informal cross-language communication is gaining momentum. As a result, the demand for translated content has reached an all-time high, and seems to be set to rise for the foreseeable future, both within Europe and across European languages, as well as beyond worldwide. From blogs to on-demand TV shows, to instant messaging and e-commerce websites, today public administrations, organisations and companies must translate large quantities of content quickly. In this section we explore how MT is currently used in our society across important areas, and we give our views on the direction in which the use of the technology could take us in the coming years.

Online MT for general use In the early stages of applying machine learning to language translation, MT was primarily used to get a rough translation of a text. Since the "MT neural revolution" in 2016, MT results have significantly improved in terms of quality, consistency, and productivity.

Nowadays, there are countless online sites that offer access to MT applications. Most belong to companies that make the systems freely available, most often with usage restrictions, to potential customers (Amazon, Google, DeepL, Linguee and Microsoft, among others); in contrast, others are provided by organisations that either facilitate MT capabilities owned by third-party companies, such as newspapers, or (often public) bodies that have their own customised system available for the general public (for instance, the European Commission and the Basque and Latvia governments, among others). Such free availability of online MT engines that can be accessed via multiple devices (e. g., computers, tablets and smartphones) makes it highly convenient for all types of users of all levels of language skills to benefit from the technology for any professional or personal need that they may have at any given time.

Currently, the free versions of systems such as Google Translate and DeepL, to mention a couple of the best-known free online services, let users translate texts between more than 100 languages by writing or copy-pasting them into a very intuitive and easy-to-use web interface, and often, even entire documents (PDF files, Word documents, Powerpoint presentation, and even entire web pages are among the supported formats) can be processed by uploading them to the platforms as easily as via drag and drop functionality. The systems even show possible alternative translations for the user to choose, and opportunities to provide feedback in the form of corrected output or preference selection. Eventually, this voluntary participation of the public allows the MT engine to improve quality.

What is interesting to see is that people use these generic translation tools to obtain translation proposals for a very diverse range of texts, which cover both professional and private spheres: from specialised texts and reports to emails, online news and school homework.

We observe that freelance, translators, professionals in non-linguistic areas and school children alike turn to the online technology to overcome linguistic barriers, to complement their language competence.

While access to these online applications is fast and straightforward, they do present risks and cultural bias. Importantly, we must remember that all text submitted for translation is genuine and often includes new information, be it personal or product- or process-related. Companies offering MT services publish legally binding notifications stating that the texts entered in the systems will be exclusively used for further system development, that is, no attempt at uncovering the information in the texts will be performed. However, it is not yet clear whether the information is completely safe, and whether some details contained in input text could somehow surface in other situations after processing. After all, we should all be weary that there are private companies behind the applications we use and that, to this day, the legal boundaries of text ownership and use are not fully regulated across Europe. Translation clients clearly requesting LSPs and freelance translators not to use freely available MT systems is a clear indication of this issue.

As mentioned, a further issue to note about these online services, and in particular those offered by big companies, is the array of languages at our disposal. It is undeniable that the collection available is increasing, with more and more languages included as research and resources allow. However, not surprisingly, it is the major and most powerful languages that benefit from the advances first and foremost, with small and minority languages repeatedly coming second-place. A quick review of the language availability clearly shows the limited provision for small languages in Europe and beyond. Not only that, the translation quality obtained for big languages is strikingly better than that available for smaller ones, showing the difference in resources and effort devoted to the different types of languages.

While the initiative to provide free MT solutions to the general public by big companies is a welcome effort, it is an undertaking to monitor closely: it cannot become a route to gathering private information; it cannot be a tool to further broaden the gap between languages, and therefore, communities; and it should be accompanied by training programs for the general public and efforts to raise awareness of the effects of using less than adequate texts.

Video Games MT has been available to the video game localisation industry for years without much success. Adding to the general resistance to trust a machine to offer translation equivalents in many languages is the need for highly creative and culturally adapted options, often with constraints dictated, for example, by available on-screen space. Localisation, including translation, is an essential part of the gaming industry and for players around the world. It enables them not only to understand the mechanics of the game and its rules, but also allows them to enjoy the gameplay and feel engaged. In other words, quality localisation enhances playability beyond mere functionality. The types of texts that need translation can vary a lot in the game industry. Besides, with the advent of online collaborative games, that have participants from across the globe partnering to fight a common enemy, uniting forces to solve a quest or simply forming communities for entertainment, in-game dialogue has become critical, and therefore, the need for instant translation from and into multiple languages.

Some game developers such as IGG, Keyword Studies and Ubisoft have their own internal localisation departments, with in-house staff and resources to address translation, among other localisation needs. Yet, there are many (game) localisation providers on the market to choose from for those who prefer an external partner to handle this aspect of the product (Alpha CRC, Altagram, BLEND, Lionbridge, Localsoft, Trágora, etc.). It is interesting to see that, although localisation companies offer the possibility to translate video games into a large number of languages (100+), the most common languages into which games are translated are the following: Chinese, German, Italian, Japanese, Korean, Portuguese, Russian, Spanish and Turkish. Yet again, this indicates an overall trend that favours big languages and enlarges the divide between major and small and minority languages.

Undoubtedly an industry on the rise, video-game localisation is challenging MT technology for instant, creative translation to allow for cross-language communication in entertainment.

Health Medical translation refers to the translation of several kinds of text, both written and oral: technical reports, medical procedures, regulatory documents, clinical histories, marketing content, etc., as well as software or training material for the pharmaceutical, medical engineering, and healthcare fields and doctor-patient and doctor-doctor communication. As in other areas, medical products and devices, documents related to clinical trials, and medical information must be available for both clinicians and patients, and therefore, presented in a language they understand.

Needless to say, this type of information is highly sensitive and requires the utmost precision and correctness given the serious consequences decisions taken based on inaccurate or incorrect information may lead to. Traditionally, medical translation has been performed by professional translators with specific subject matter knowledge, able to deal with the sensitive technical and regulated nature of medical documentation and to guarantee high translation quality.

In this context, MT has been widely discarded for years, as it was perceived as a tool which could not guarantee the level of accuracy required in terms of terminology but also in meaning transfer in general. If this field is to benefit from MT, there is still work to do, in particular, with regard to meaning transfer accuracy. Engines which can provide translations good enough for gisting or that are of moderate quality do not suffice any more. It is time to push for MT accuracy and consistency, and accept nothing short of high quality translation (Haddow et al., 2021).

Aside from written text translation, MT could prove of great assistance in this field in doctor-patient communication contexts. The increase in migratory movements results in medical services treating immigrants with severe communicative difficulties. It is at this time when a medical translation plan to help communication between patients and doctors becomes of great importance. While professionally trained medical interpreters remain the go-to specialists for translation in clinical practice and well-resourced institutional settings, they are often not available or provided by the healthcare institutions or covered by insurance companies. For this reason, physicians are beginning to turn to MT to facilitate communication, with all the risks that are involved. MT might be used to clarify patient histories, review a clinical diagnosis, or convey recommended treatment plans and follow-ups to facilitate comprehension. Physicians might also encourage patients to ask questions or respond to queries by directing them to input text into the machine translator. Once again, the system must be ready to deal with sensitive data in a very specific domain.

What is even more critical than in other fields is that, often, the language pairs involved in this type of communication involve non-mainstream languages, which vary from location to location according to the migratory flows the regions receive. If MT is to assist in this context, therefore, better systems that can specifically tackle the local languages and those of the immigrants are essential.[19]

MT has been an useful tool to diminish borders between people and the COVID-19-related information.[20] To make emergency and crisis-related content available in as many languages as possible, the Translators without Borders together with several academic (Carnegie Mel-

[19] Even today MT solutions provided by global companies support more commonly spoken languages, e.g., Google Translate and Microsoft Translator are only available in 108 languages, while Amazon Translate supports only 75.

[20] For example, in 2020 SDL made available SDL Machine Translationto all organizations, researchers and professionals engaged in any aspect of COVID-19 medical research, discovery and development; more information at https://www.biospace.com/article/releases/sdl-offers-machine-translation-free-of-charge-to-health-science-professionals-/, while NAVER LABS Europe released a MT model for Covid-19 research https://europe.naverlabs.com/blog/a-machine-translation-model-for-covid-19-research/.

lon University, George Mason University, Johns Hopkins University) and industry (Amazon, Appen, Facebook, Google, Microsoft, Translated) partners have partnered to prepare COVID-19 materials for training state-of-the-art Machine Translation (MT) models for nearly 90 languages.[21] The COVID-19 Multilingual Information Access initiative[22], a collective effort from the LT community, aims to improve information exchange about the virus across all EU languages and more by accelerating the creation of resources and tool. However, even in case of COVID-19-related information MT also come with limits (for example, language coverage, terminology or technical context, as well as literary texts) where human interpretation is necessary.

Public Administration and Legal Documentation Making legal and administrative documents available in the (at least official) languages of the different regions across Europe is an obligation of national governments. Given the intricacies of this type of texts, which require full accuracy, careful use of terminology and often, the convergence of different legal systems, translation has largely remained in the hands of human translators. However, the large amount of public documents available to train new systems and the techniques to guide terminology selection in automatic systems is little by little allowing MT to enter the field. Several projects, actions and initiatives, for example, PRINCIPLE (Way et al., 2020), ELRI[23], ELRC[24], aim to prepare and share language resources that can improve translation services.

Along the same lines, the availability of high-quality neural MT across many language pairs capable of being used intensively at different levels of public bodies, Member State and Public Administration has been put forward as a key priority for the European Commission, particularly for under-resourced EU languages and in some priority domains. For example, NTEU project[25] aims to create and release near-human quality machine translation engines for public administrations (Bié et al., 2020), while the iADAATPA project (Castilho et al., 2019) developed innovative platform to offer state-of-the-art domain-adapted MT engines to public administrations (Castilho et al., 2019).

As stated by the European Language Resource Coordination[26], an excellent example of how EU Council Presidency staff members and public administration translators can utilise AI-powered NMT to facilitate and speed up multilingual communication is demonstrated by Finland. During its 2019 Presidency of the EU Council, Finland set a record for the EU Presidency Translator usage: more than 740 thousand translations containing over 12.7 million words. By combining the CEF eTranslation service with custom NMT engines developed by Tilde, the EU Council Presidency Translator was used to translate text snippets, documents, and websites using a responsive online translation website and a CAT tool plugin (Pinnis et al., 2020) . The main users for the translation tools included EU Council Presidency staff members, public sector translators in the hosting country of the Presidency, EU delegates, and international journalists covering the events. On top of the custom NMT systems, the EU Council Presidency Translator provides access to all the MT systems from the CEF eTranslation service, for translation between the 24 official languages of the EU and English.

The challenge to address here is the provision of this type of service not only to the 24 official languages, but to all languages in Europe, promoting citizen equality and European cohesion, key to a stable and unified view on the region. The rapid development opportunities translation technology offers and the techniques for low-resource languages that are being explored coupled with the identification of key information relevant for EU citizens, could be used to present Europeans with significant pertinent documentation in their lan-

[21] https://tico-19.github.io
[22] http://eval.covid19-mlia.eu
[23] http://www.elri-project.eu
[24] https://www.lr-coordination.eu
[25] https://nteu.eu
[26] https://www.lr-coordination.eu/index.php/news/AI-powered-language-technology-helping-to-shape

guage of preference.

eCommerce To fully reap the rewards of the international markets, every piece of content on an e-commerce website should be translated into the target customer's language. This increases the (potential) customer's understanding of the features of the product but also helps build trust and the feeling that the company cares about its clients.

eCommerce companies around the world face the challenge of efficiently and continuously translating content across diverse devices and channels. The sheer quantity of eCommerce content and its constantly changing nature, make it an ideal beneficiary of MT. Product information is the most obvious element that requires translation, but the same is true for peripherals such as opinion messages, blogs, social media, marketing messages, and so on. It is clear that eCommerce requires a mix of technical, highly accurate translations and informal, creative, culturally aware translations.

Today there are many companies (e. g., Lionbridge, Protranslating, SeproTec, Simultrans, Smartling, Stepes, Supertext, Weglot) that can help online business owners to make their content multilingual to reach different markets and potential new customers. The services offered by these companies range from the automatic translation of the vendor's website to the integration of different MT engines through APIs on CAT tools used by human translators and on chats for customer support. What is interesting is that many of them offer plugins compatible with the most common CMS and eCommerce solutions in the market, such as WordPress, Drupal, Joomla, Magento and WooCommerce, allowing for seamless translation workflows.

The quick review through some of the key consumer sectors underpinned by ICT shows that a seamless integration of MT services could greatly broaden their reach and facilitate positive experiences by enhancing them via use of native languages. For this to become a reality, MT research and implementation need a significant push. The current shortcomings of the technology and areas where effort should concentrate revolve around aspects that help increase trust at times through increased accuracy, sometimes through high cultural adaptation and creativity. It is high time MT quality and suitability are accounted for not only by means of usage-agnostic metrics, but also by customer experience measurements. It is clear that a region where all citizens feel equal and with the same quality access to resources, services and commerce will do nothing but boost European cohesion.

5. Machine Translation: Main Gaps

5.1. Data

5.1.1. Availability

As stated in the EU Charter and the Treaty on the EU, all 24 official languages of the EU are granted equal status. Despite multilingualism being a core tenet of the EU, the META-NET White Paper Series found that 21 of the 30 European languages investigated were at risk of digital extinction, with no language being considered to have "excellent support", only English was assessed as having "good support" (Rehm and Uszkoreit, 2013). In addition to the official languages, there are over 60 regional and minority languages, as well as migrant languages and sign languages, spoken by 40 to 50 million people. The negative consequences of this lack of resources are twofold. Not only are Europeans not receiving the digital resources they are entitled to, but this lack of resources also represents a lack of language data, which can be used to train MT engines. As the META-NET findings show, availability of language data needs to be tackled.

However, availability does not guarantee usability. The Open Data Directive (2019/1024/EU) does not recognise language data as a high-value data category, meaning what little language

data does exist for at-risk languages may not necessarily be clear, or, how data fit for re-use for the purposes of developing MT systems to support these underrepresented languages.

5.1.2. Usability

To be considered usable as training data, language data must meet certain criteria. For instance, to train high-performance neural MT systems, bilingual data needs to be correctly aligned (both from a structural and from a content point of view), it needs to be adequately cleaned, and it needs to undergo a number of necessary pre-processing steps (e.g., tokenisation, byte-pair encoding, etc.). While optimised automated data cleaners can tidy up any bilingual data to a satisfactory point, translation data in the form of TMX and TBX files is an ideal starting point for building MT systems.

The ELRC White Paper *Why Language Data Matters* found that specialised user training and negative dispositions towards CAT tools (along with their high costs) were blocking factors for translators to embrace CAT tools, hindering the creation of TMX and TBX files and language data sharing (ELRC, 2019).

The collection of usable language data in Europe is particularly important: while the intensive use of popular systems developed by large global MT service providers allows them to collect and re-use user data, services in Europe would not be able to re-use user data in this way due to GDPR (Aldabe et al., 2021).

5.1.3. Domains

Neural MT systems benefit from exposure to a wide variety of data, including style and content variety. Likewise, while domain specificity is important to tune an engine towards a particular field or subfield, expanding the domain coverage usually brings benefits to the training of a neural MT system. This means that domain availability is almost as relevant as language availability. While data categories such as legal, financial, and technical are usually well covered in terms of availability and suitability of training data, more specific or uncommon sub-categories may not have comparable amounts of training data available. Moreover, there is generally a disparity between publicly available and proprietary bilingual corpora in terms of domain coverage. As a result, there is a gap in the availability of subdomain-specific language data both in official and minority languages. The danger of this gap is that it could lead to the centralisation of some specialised fields over others, excluding speakers of less supported languages in the long term.

5.2. Technology

5.2.1. The "Compute Divide"

With the paradigm shift to neural MT systems, research and development related to MT have become increasingly computationally intensive. It follows that access to hardware, experts, and involvement in research have also shifted in such a way that elite universities and large firms have an advantage due to their ease of access to computing power (Ahmed and Wahed, 2020).

According to the ELE report on existing strategic documents and projects in LT/AI, there is a lack of necessary resources (experts, High Performance Computing (HPC) capabilities, etc.) in Europe compared to large US and Chinese IT corporations (e.g., Google, OpenAI, Facebook, Baidu, etc.) that lead the development of new LT systems. The report also highlights an uneven distribution of resources, including scientists, experts, computing facilities, and IT companies, across countries, regions and languages (Aldabe et al., 2021).

5.2.2. Multi-modal MT

MT is commonly thought of as translating text to text by means of computers. Multi-modal MT is possible, but it is still relatively in its early stages. Fields in which further technological innovation would increase potential use cases for MT include image recognition, text-to-speech and speech-to-text.

Image-to-text translation makes use of Optical Character Recognition (OCR) to isolate text in images. This technology is quite effective, and nowadays smart phone and tablet users can generally avail of image translation services free of charge. However, OCR software is not as widespread as standard text-to-text translation. Multiple factors affect the accuracy of OCR technology, including coloured or decorative backgrounds, blurred texts, and skewed documents. It can also struggle with non-Latin alphabets, larger or smaller letters, look-alike characters, and handwritten text (Dilmegani, 2020). All of these issues affect the accuracy of OCR and can result in errors. In MT applications, these errors may result in nonsensical translations. Combining OCR with text prediction is an open area of research and may improve the accuracy of this technology.

Audiovisual media plays a more and more central role in our lives, for instance when thinking about the increasing popularity of AI-powered virtual assistants and online streaming services. For this reason, the ever-growing demand for translation of audiovisual content has sparked interest in the development of MT-centric text-to-speech and speech-to-text applications. Moreover, the need for accessible content in the form of subtitles and audio descriptions for those who are visually impaired, deaf, or hard of hearing has the potential to drive innovation in MT.

The New European Media Strategic Research Agenda states that in the future AI will be used to translate speech to subtitles, text to Sign Language and Sign Language to text (New European Media Initiative, 2020). Its recommendations include:

- Fluidising/streamlining the circulation of audiovisual (or video) programs through MT, while humans focus on the quality of work, for example.
- Encouraging synergies and convergence between subtitling and the development of multilingualism or the integration of foreigners (migrants for example).
- Developing AI tools for automatic translation from speech to subtitles, from text to Sign Language, and from Sign Language to text.
- Developing AI tools for robust automatic translation of subtitles (multi-languages).

Training high-performance MT systems to translate subtitles is particularly challenging. As mentioned above, rigid copyright laws in Europe forbid the use of translations of copyrighted movies and audiovisual material, despite the fact that this may constitute fair use (see section 5.4.1). Compared to technical language, subtitles are often more creative and idiomatic in nature, increasing the difficulty of translation and the need for high volumes of good-quality training data.

5.3. General Approaches

5.3.1. Project Manager as end user

Project managers in the language industry are often faced with pressure to provide discounts when using MT under the premise that MT boosts productivity, allowing linguists to post-edit more words per hour than if they were to translate from scratch. While the advent of MT has allowed translators and linguists to spend less time on repetitive content, productivity gains still depend on several other factors, including the quality of the MT systems and the

complexity of the content or domain. The pricing pressure often arises as a consequence of not taking into account these extra factors which make MT post-editing a more complex task than someone unfamiliar with MT might initially think. Providing project managers with the resources to better communicate these factors could be a step towards relieving pricing pressure.

5.3.2. Linguist as end user

LT has changed the role of the translator, here we use the word *linguist* to refer to language professionals who translate, post-edit, and evaluate LT among other tasks. As the ELRC report *Why Language Data Matters* has shown, some linguists express negative dispositions towards MT and CAT Tools. This has slowed down the adoption of LT along with the creation of quality language data:

> "For instance, in Croatia and Slovakia, little awareness of the availability of technical solutions that could improve the translation process was demonstrated as well as a generally slow adoption of technologies. In other countries (e.g., Ireland, Greece, and Romania), an unskilled use or non-use of CAT tools was observed. The Irish Technology National Anchor Point (T-NAP) held a training workshop on the use of CAT tools and MT post-editing in June 2019. Less than 20% of the translators (public servants and freelance translators) had used CAT tools before, clearly indicating the big demand for this kind of training." (ELRC, 2019)

There tends to be a generational divide in attitudes towards the adoption of MT in translation workflows among linguists, with some older linguists fearing that MT threatens their job security. Younger linguists tend to have more positive dispositions due to proper training in such technologies being included in their higher education courses. Moreover, they will usually have been introduced to MT at a stage where it was much more reliable than the rule-based or statistical systems that older linguists would have used in their training.

However, linguists play an important role in the assessment and continuous improvement of MT engines, especially because there is still no universal way of automatically evaluating MT quality (see section 3.4). For this reason, while the role of traditional translators might have changed with the overall improvements in MT, demand for linguists has remained high alongside the developments of AI and LT.

5.3.3. Reader as end user

At the other end of the spectrum, hype about the advancements of AI and MT can guide those less familiar with LT expertise into thinking that MT is infallible (see Läubli et al. (2018) and Toral et al. (2018) for clear demonstrations that the 'human parity' claims were less than watertight). The language industry has seen a growing interest in MT Literacy, where linguists less familiar with LT are taught to use MT critically (Bowker, 2021).

The wide availability of MT applications coupled with the sometimes-deceptive fluency of neural MT output may lead users to avail of MT uncritically, without always understanding its pitfalls. As more and more non-language professionals employ MT daily, there is a growing need for adequate literacy which includes a measured understanding of LT and its capabilities among the general public. One step in this direction includes educational publications such as *MultiTraiNMT: Machine Translation for Multilingual Citizens*, an open-access e-book which addresses the technical foundations of machine learning as used in MT and the ethical, societal, and professional implications of its use (Kenny, in preparation). Work demonstrating what MT can and cannot do in the area of literary translation is also worthy of consideration (e. g., Moorkens et al. (2018); Toral and Way (2018)), as is recent work on 'translationese' (e. g., Toral (2019); Vanmassenhove et al. (2019, 2021)).

5.3.4. Automated Evaluation of MT

Automated metrics are a language-independent and cost-effective way of assessing the quality of machine translated output. Research in the field focuses heavily on developing metrics that are able to show higher and higher correlations with human judgement, and as a result, several different metrics are presented at conferences around the world every year. Despite, or perhaps as a result of, their abundance, there is still a lack of agreement among the MT community on a single metric which can be used universally to assess the quality of MT engines prior to deployment.

Bilingual Evaluation Understudy (BLEU: Papineni et al. (2002)), for example, has enjoyed perhaps the broadest use in the MT industry, despite its known shortcomings with regards to neural MT. Many other metrics have been developed since BLEU, and while they all have their pros and cons, the widespread use of BLEU has proven that metrics can serve a purpose without being scientifically infallible. Adopting a single metric as a standard for measuring NMT would possibly allow for a widespread benchmarking of LT across Europe.

5.4. Regulation

5.4.1. Licensing and Copyright

Translation memory and terminology data is often licensed for non-commercial use only. When commercial licences do exist, their prices are often prohibitively high. This acts as a major barrier to SMEs developing MT applications, especially when there is a limited amount of data available.

With regards to copyrighted content, copyright laws pose a further barrier in Europe. While copyright law is subject to fair-use exceptions in countries such as the US, European law is far less flexible. Many European laws severely restrict the use of parts of copyright works for purposes such as data mining. However, using subtitles from a copyrighted film to train an MT engine would not necessarily detract from the profits expected from the owner, as users would have no access to the engine's training catalogue. If lawmakers could agree that using aligned translations of copyrighted data constitutes fair use, as far as it in no way impairs the value of the materials and does not curtail the profits reasonably expected by the owner, LT stakeholders could avail of this high-quality language data.

An important new book on this topic is Moniz and Escartín (2022), and many of the chapters in that volume are relevant for a wide variety of issues on this topic.

5.4.2. Legislative and Adoption Gaps

Despite the widespread celebration of multilingualism in the EU, there is no common policy addressing language barriers as of yet. Below are some examples of scenarios where multilingualism acts as a barrier to people in times of crisis. It is fair to say that current legislation does not account for these scenarios, resulting in critical gaps in services for communities in the EU. Adopting MT in these areas could mitigate the difficulty sometimes caused by language barriers, strengthening the position of multilingualism as a facet of European identity that is worthy of celebration.

The Covid-19 pandemic has demonstrated the need for rapid dissemination of clear information and guidelines in times of crisis. In Ireland, the provision of multilingual information was seen to be slow, reactive, and random. Even the provision of information in Irish and Irish Sign Language was slow in the early stages (O'Brien et al., 2021). The first recommendation made in a report funded by the Dublin City University Educational Trust, *Communicating Covid: Translation and Trust in Ireland's Response to the Pandemic*, is for state departments to implement a coordinated approach to the provision of translated content in crises. Careful

use of MT can increase productivity in times of crisis, ensuring that public health information is communicated quickly and effectively. Ensuring that language professionals trained in using MT do so before multilingual guidelines are proofread and published ensures that members of the public, who might not use MT critically, have their language needs met.

The requirement for all translations of personal documents, including birth, death, immigration, adoption certificates etc. to be stamped by a sworn translator can exacerbate particularly stressful times in the lives of civilians, by adding costs and waiting times. The repetitive nature of documents like these as well as their standardised terminology are particularly well-suited to MT.

Just as the Audiovisual Media Services Directive boosted demand for text-to-speech and speech-to-text technologies, there could be an increase in the demand for MT if policies necessitating the translation of certain audiovisual material into all 24 official languages were introduced. While EU Law requires that the product descriptions of goods sold within the EU be translated into the Member State's official language, as of yet there are no such regulations regarding product descriptions for cross border e-commerce.

Finally, there is a gap in publicly available MT services which cater specifically to the needs of people in Europe. With the increased use of smartphones, MT applications have become a feature of daily life even for non-language professionals. Users around the world avail of free-of-charge MT services provided by Google, Microsoft, Baidu, etc. However, these major global companies could start withdrawing or charging for their services at any time. Moreover, they do not cater specifically for the needs of European citizens.

Training neural MT engines is resource intensive and has a heavy carbon footprint. One area where the law is perhaps too relaxed is in relation to carbon emissions in the field of AI research and development. Researchers have warned of the marginal performance gains associated with expensive compute time and non-trivial carbon emissions. Strubell et al. (2019a) recommend that time spent retraining should be reported for NLP learning models and that researchers should prioritise developing efficient models and hardware. The EU has the opportunity to be a pioneer in training and developing green LT by following and enforcing these recommendations.

6. Machine Translation: Contribution to Digital Language Equality and Impact on Society

Nowadays, due to globalization, MT is essential for the development of society. It impacts directly on the economy and cultural exchange between countries. Human translators cannot meet the huge demand for translations in a limited time at any time and also at a low cost. Translations must be quick and only MT can work in this scenario. Otherwise, translations are not helpful. Anyone can access MT allowing for the democratisation of information in any language. For human translators, MT is a useful tool to facilitate their work. It is much faster and requires less effort to post-edit MT than translating from scratch.

6.1. History and Background

Since the MT hype prior to the ALPAC report in the 1950's MT was hailed as the technology that could "translate all the Russians do in one week". The focus was then on information services. The concept of well-resourced languages did not exist as MT depended on large sets of rules between close or distant language pairs – and with it a varying degree of success and accuracy. The interest in MT in the days of rule-based systems focused on large language pairs, or at least language pairs with the greatest demand for translation services and documentation. Languages which lacked a sufficiently large number of speakers to justify

demand for development or lacked sufficient political interest began to suffer a "technology gap" because of a lack of interest in development. Thus, languages such as French, Russian, German, Spanish, Italian, Japanese, Chinese and Arabic were the first to be developed from English and sometimes among themselves. It goes without saying that the success was much greater if the languages were syntactically and semantically closer than, let's say Japanese-English or French-Arabic.

In 1984, example-based MT appeared. With these systems, translation memories were started to be collected and the translation examples were used to match parts of the new translations according to the phrases.

With the advent of statistical systems from the early 2000's, the focus shifted to having as much available data as possible. Many resources and in particular massive amounts of parallel data are required to build solid MT systems. It is from this time that we started to witness the concept of a "digital gap" between well-resourced languages that can bank on several kinds of resources, from national libraries to a long tradition in literature and translation to other languages – both in Europe and elsewhere – which lack the sufficient digital footprint and resources to keep up with the requirements of massive amounts of data that are needed to build MT models.

The EU's DGT and the Europarl corpus have been made available to academia and industry for a long time as repositories on which some base systems could be built. However, in the case of countries that joined the EU after the large enlargement, there is a lack of parallel data because of the shorter amount of time those countries have been official EU languages. Parallel data creation is costly in terms of time and resources and many times dependant on works of Public Administrations. The NEC TM project[27], for example, calculated in its market study that European Public Administrations spend approximately €300 million a year in translation services contracts with language vendors. This parallel data is mostly not requested back by institutions – many of which operate in low-resourced languages.

Data availability, in short, goes hand in hand with the availability of MT and MT quality and the contribution it can make to DLE and of course its impact on society. There is a strong link between data pipelines to improve local (national) technology, the awareness that citizens are also data producers (and Public Administrations are always the data creation champions) and the creation of MT technologies. For example, in the case of Catalan, a regional language in Spain that is co-official in 3 Spanish regions but it is also spoken in parts of France and Sardinia (Italy), having co-official status kickstarted a series of administrative decisions that facilitated the creation of more and more parallel data, which has been utilised as a basis for local MT companies. Prior to it, some of those companies and Universities (U. Alacant) were involved in the creation of rule-based systems (Apertium) as a result, then turning and supporting ML companies such as Prompsit or Pangeanic.

Societies that care about data sovereignty and establish language data policies can facilitate the growth of LT companies. The question now is if these companies and technologies in turn create a virtuous circle, impacting society.

6.2. Does MT contribute to Digital Language Equality?

The availability of well-known online MT services has revolutionised the translation industry. In its wake, it has also affected expectations about language translation availability, ubiquity, quality and embedment as part of many services. Large technology companies began rolling out sets of languages (mostly into and out of English) as soon as statistical MT systems became stable and could scale. Of course, well-resourced languages were the first to be served, but such systems soon made available high-quality systems in over 100 languages,

[27] https://www.nec-tm.eu

with the added advantage of better perceived quality as neural engines began to become the state of the art from 2017/18.

Many of these systems have been built using a non-negligible amount of back-translation, a technique through which a human-quality source language is translated into English (basically flipping over the order in which the encoding and decoding will happen). This and other techniques such as data augmentation, have made it possible to scale in the data creation of parallel corpora and, consequently, of more and more MT systems even in low-resourced languages. The key point is if all those systems by large IT companies and the increasing number of systems by smaller players contribute to a more equal relationship between languages from a digital.

One would be inclined to acknowledge the paramount role of MT in knowledge acquisition and information retrieval from any language, once a system is built. It allows data from those low-resourced languages to be understood and be made digitally available to much wider audiences and the world's public. It gives access to previously relatively unknown information, as it comes in a language with small number of speakers or low resources to be made available quickly and vice versa: those same speakers can access information in any of the better-resourced languages. In short, new developments in MT do bridge the gap in information access to "minority", "low-resourced" languages or languages lacking enough digital footprint.

Transfer learning techniques have been proved to help low-resource languages training multilingual models. As mentioned in Yang et al. (2021), one technique is using huge pretrained language models to initialise the models. In 2021, multilingual models have beaten bilingual models in the MT competition in the Workshop of Machine Translation (WMT) in many language pairs by Facebook submissions Tran et al. (2021a) including Hausa which is a low-resource language spoken in West Africa.

6.3. Uses of MT in Society

Uses of MT are very varied, from customer reviews on travel sites to document translation for tax o legal public administrations. None of those uses and the business intelligence that can be derived from them can happen without language transfer. MT not only works for equality on dispute resolution or as a source of information for insights at scale irrespective of the source, but also enables business to build on those services, impacting the society they belong to. We cannot separate the use and availability of the technology from its societal impact.

The ubiquity of MT services is an indisputable fact of current European digital societies. It has become an embedded technology in many services and it is expected as a service, real-time and of high quality. The ELE consortium has identified several day-to-day uses which serve as an example of how it is used in very different spheres of our lives.

- Politicians verify national legislation of other EU Member States by machine translating it to compare to their own and /or gain inspiration from what is being legislated in other countries.
- Citizens communicate with free online tools during their visits to other countries, or to check text (tourism and general purposes).
- The general public use free online tools to understand instructions, documents, sayings, news, etc.
- Social media conversations from sports events to political issues. Most happen in a monolingual setting but information sharing does happen in platforms such as twitter, and about specific topics; there have been studies on the radically different views of

the Euro crisis in Germany and Greece, for example, with public opinion being inflated as stereotypes were reinforced but serious information sites – which were machine translated – offered a more in-depth view from both sides.

- Students machine translate papers, websites, research from educational establishments all over Europe.
- E-commerce websites offer products online to consumers within the Union in multiple languages.
- Public Administrations translate documentation when text and documents from another public administration when the need to liaise and exchange information. This applies to tax records, justice, medical records, etc.
- Translation companies many times use MT services for daily work in order to provide faster and cheaper services to their clients.

All these uses generate massive amounts of online data, that is not re-used or it is generated for the benefit of the free online tools providers to make their engines and technology more accurate. Access to massive amounts of data that is freely available and provider by general users has scaled a lot of MT research, whilst it has provided little in terms of open-sourced, generally available resources.

Whilst the majority of the talent in development has been European, large-scale developments are foreign to the EU or the result of private sponsorships. Heavy investment in MT research at Universities over the years created the know-how and technical knowledge which has not been exploited commercially.

The question for Europeans remains on privacy of the data used and how this data is transmitted. The MT landscape is dominated by large non-European players and technology companies such as Google Translate, Bing, Yandex. DeepL is the only EU-based provider, being sponsored by a German initiative born as a result of parallel text data collection over many years (Linguee). Most European MT companies remain fairly small and have little impact (visibility) on society beyond professional level usage. The EU's own service (eTranslation) is available for free to public administrations and it also opened its services to SME in 2021, although connection is by means of API only (no panel), putting it on an equal footing with other European commercial solutions.

A good example of increasing privacy concerns and a good lesson to decision-makers comes from Switzerland, where DeepL was recently banned by management as an external tool amid concerns of MT privacy and data exploitation. Swiss Post declared as policy that its staff should only use its own MT technology, so no private data or data belonging to the organisation would be sent to third parties.[28]

GDPR has the potential to change things as privacy concerns become relevant to institutions and enterprise, with EU-sponsored projects such as MAPA[29] providing highly accurate, open-source anonymisation for public administrations. It remains to be seen how this potential is exploited so that MT and general NLP solutions permeate and help create a more data-based Europe based on intelligent solutions with the citizen at its core.

[28] https://slator.com/swiss-post-bans-deepl-backs-down-after-staff-uproar/
[29] https://mapa-project.eu

7. Machine Translation: Main Breakthroughs Needed

7.1. Rationale

The needed breakthroughs we list here are inspired by the following observations: (1) there is a need to investigate new ways of developing MT systems on future-proof hardware (2) EU policies encourage more fundamental research into carbon neutral and trustworthy AI (3) although policies for data and AI are in place, data collection remains a struggle for all parties involved in MT development (4) a great deal of groundwork on novel technologies still needs to be done.

7.1.1. The need for new hardware infrastructure and training paradigms

According to a competitiveness analysis ordered by the Connecting Europe Facility (Vasiljevs et al., 2019), the position of the European MT market, as compared to that of North America and Asia, is excellent for research and innovation, while it lags on investments, infrastructure and industry implementation. At the same time, the study highlights that the market is fragmented (STOA, 2017, p. 104):

> "The European HLT industry is mainly made up by innovative smaller companies and micro-enterprises. Although most of them have been established in the market for several years, the specificities of the LT market in Europe (local/national companies with expertise in local languages that serve local markets) hamper their growth. The transformation into global players capable of competing with non-European big companies requires financing in all stages of business life cycle, not only in research activities. The creation of investment instruments and accelerator programs can increase the economic potential of the high level of innovation within the European industry, which could remain European instead of becoming subsidiaries of the USA's big players."

This fragmentation causes serious issues for the level of intensity at which LT research can be conducted. While in North America and Asia resources can be allocated to only a limited number of languages, in Europe, resources must be distributed across a multitude of official and unofficial EU languages. As a result, the scale at which European research can be conducted is limited. Considering the massive infrastructure that is required to train very large state-of-the-art LT systems, Europe starts with a systemic handicap. When looking forward to 2030, we expect the movement towards more efficient and real-time translation to continue[30]. Europe's strong foundation in research and innovation can compensate for the disadvantage European organisations have with respect to infrastructure, provided that a concerted effort is undertaken in researching the development of new hardware platforms and respective AI training paradigms. Hence, a breakthrough in these fields is needed for Europe to remain on par with the rest of the world.

7.1.2. Alignment of needed breakthroughs with existing EU policies

The breakthroughs in the development of hardware platforms and training paradigms are also warranted by several EU policies.

[30] Google reportedly uses AI to accelerate the design for AI chips that can be used for faster and less power-consuming MT, as described in Mirhoseini et al. (2021)

Through the European Green Deal[31] and the Horizon Europe Work Programme,[32] the European Commission has committed to making "Europe the world's first climate-neutral continent by 2050". To achieve this, the economy must be transformed with the aim of climate neutrality. More efficient AI infrastructure can help in reducing the amounts of energy that are required for data storage and algorithm training[33]. For example, an MIT study (Strubell et al. (2019b)) found that training a large AI model to handle human language can lead to emissions of nearly 300,000 kilograms of carbon dioxide equivalent, about five times the emissions of the average car in the US, including its manufacture. In line with this study, Swedish researchers have forecast that data centres could account for 10% of total electricity use by 2025.[34]

An equally fundamental breakthrough is needed in the understanding of how our current algorithms work. It is rightfully observed that "recent Natural Language Processing (NLP) systems based on pre-trained representations from Transformer language models, such as BERT and XLM-Roberta, have achieved outstanding results in a variety of tasks. This boost in performance, however, comes at the cost of efficiency and interpretability. Interpretability is a major concern in modern Artificial Intelligence (AI) and NLP research, as black-box models undermine users' trust in new technologies."[35] The EU Coordinated Plan on Artificial Intelligence[36] recognises this problem and advocates the need for trustworthy AI, mainly from the perspective of the end-user. But interpretability and explainability of AI models are also of great importance for the scientific community. If researchers wish to improve their algorithms, they must gain a deeper understanding of what causes models to behave the way they do. Introspection is needed to prevent models from performing poorly or to act in a gender or culturally biased manner.

7.1.3. Policy breakthroughs needed in support of AI development

The EU Coordinated Plan on Artificial Intelligence correctly states that "Further developments in AI require a well-functioning data ecosystem built on trust, data availability and infrastructure."[37] But it underestimates the effect that one of its cornerstones has had on data collection in the field. According to the plan "The General Data Protection Regulation (GDPR) is the anchor of trust in the single market for data. It has established a new global standard with a strong focus on the rights of individuals, reflecting European values, and is an important element of ensuring trust in AI. This trust is especially important when it comes to the processing of healthcare data for applications driven by AI. The Commission would like to encourage the European Data Protection Board to develop guidelines on the issue of the processing of personal data in the context of research. This will facilitate the development of large cross-country research datasets that can be used for AI."[38]

Unfortunately, the GDPR has had an adverse effect on a large part of the European LT industry. Stakeholders in data management, publication and collection have come to incorrectly assume that all data is personal by default, as an overly cautious measure to comply with the GDPR. This is especially true for data consisting of human natural language, which is difficult to manage, since it has no fixed schema indicating when personal details (for example names or addresses of persons) may occur. As a result, expensive legal counsel and

[31] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52019DC0640
[32] Horizon Europe Work Programme 2021-2022, European Commission Decision C(2021)4200 of 15 June 2021
[33] See for example https://ec.europa.eu/research-and-innovation/en/horizon-magazine/ai-can-help-us-fight-climate-change-it-has-energy-problem-too
[34] See for example Andrae (2017)
[35] See for example https://eval4nlp.github.io/sharedtask.html
[36] Coordinated Plan on Artificial Intelligence, COM(2018) 795 final
[37] Ibid.
[38] Ibid.

tools for anonymisation are applied in situations where they could be avoided or are not necessary at all. In addition, non-European AI companies have been able to operate without GDPR restrictions ever since, which has gained them a considerable competitive advantage over EU companies.

Although the EU Coordinated Plan on Artificial Intelligence has foreseen a framework for the free flow of non-personal data in the European Union[39], including the creation of common European data spaces in a number of areas, and a proposal for a directive on the re-use of public sector information[40], the process of obtaining linguistic data that has been created using public funding is currently too cumbersome and pull-oriented. This means that data resulting from public procurement procedures has the tendency to remain locked up in privately-owned data silos, while the research community and LT industry must go to great lengths to find, identify and reconstruct the public part of this data using imperfect NLP tools (see for example Koehn et al. (2005)). Instead, a crucial breakthrough could be achieved if existing policy frameworks were adapted to make it mandatory for Member States to make all data in natural language-related workflows publicly available. It is the LT industry's mission to reconstruct human thought processes in an automated way. Human operations on linguistic data such as translation, revision and correction of translations, summarisation, simplification, etc. can provide the necessary data points to train AI algorithms to achieve this mission. A policy-inspired push model would be greatly beneficial for the development of all related research domains. As a first step, public services administrators should be made aware of the value of their human workflows. As a second step, the intellectual property resulting from public services workflows should be released to the public domain by default. As a third and final step, workflow data should be made discoverable in a pub/sub (publication/subscription) manner, so it can be picked up easily by interested parties.

7.1.4. Realism

In what follows, we list a couple of needed breakthroughs that may be within reach by 2030. To ensure that policies can be developed to foster advances in the selected topics, we only included topics that tick all SMART boxes, i. e. topics that are specific, for which the results are measurable, attainable, realistic, potentially deliverable within a timeframe of ten years.

7.2. Hardware/Software Codesign

If we want MT to become ubiquitous, especially in embedded devices, the hardware on which it runs must be scaled down. Currently, large hardware infrastructures are required to accommodate for the required computation power and storage of Deep Neural Nets. Consequently, the models that run on such hardware must be adapted accordingly. Such adaptation must occur with a minimal loss of quality, while increasing translation speed and reducing power consumption. To achieve this, a breakthrough in MT hardware and software codesign is required. Note that both must be developed in cooperation, to ensure that the capabilities of the hardware are aligned with the needs of MT training and inference.

In a survey paper on the topic Dhouibi et al. (2021) concisely describes the problem: "Deploying [...] Deep Neural Networks (DNN) on embedded devices is still a challenging task considering the massive requirement of computation and storage. Given that the number of operations and parameters increases with the complexity of the model architecture, the performance will strongly depend on the hardware target resources and basically the memory

[39] Regulation (EU) 2018/1807 of the European Parliament and of the Council of 14 November 2018 on a framework for the free flow of non-personal data in the European Union

[40] European Parliament and Council Directive on the re-use of public sector information, COM(2018) 234 final

footprint of the accelerator. Recent research studies have discussed the benefit of implementing some complex DL applications based on different models and platforms. However, it is necessary to guarantee the best performance when designing hardware accelerators for DL applications to run at full speed, despite the constraints of low power, high accuracy and throughput."

Several approaches to replace GPU-based computing are currently under investigation: "The training and the inference phases of DL models are being executed on powerful computation machines using advanced technologies such as new multicore Central Processing Unit (CPU), Graphics Processing Unit (GPU) or clusters of CPUs and GPUs. Usually, GPU platforms are better on supporting training and inference of more sophisticated models. GPU technology offers a high computation capacity but ensures the interdependence of the data which is expensive in terms of power. Application Specific Integrated Circuits (ASICs) can achieve even higher performance and can improve the energy efficiency, which is a key factor in embedded systems. However, the deployment of DL model on a customised ASIC requires high investments due to a long and complex design cycle. Recently, FPGAs have become a promising solution to accelerate inference, they offer the performance advantages of reconfigurable logics with a high degree of flexibility. Specific hardware design on such platforms could be more efficient in speed and energy compared to other platforms. Moreover, the deployment of large-scale DNNs with large numbers of parameters is still a daunting task, because the large dimensionality of such models increases the computation and data movement. So, to deploy such sophisticated models in embedded platforms and to obtain a more robust model, the internal operations and number of parameters can be reduced by optimising the network architecture. Several optimisations techniques were discussed in the literature. One of the most popular optimisation approaches that makes models faster, energy efficient and more hardware friendly is model compression, which includes the low data precision, pruning network, low-rank approximation, etc. Furthermore, for efficient implementation of an optimised DL model, further acceleration improvement is required. Indeed, it is necessary to maximise the utilisation of all offered opportunities at several levels of hardware/software codesign to achieve high performance in terms of precision, energy consumption and throughput." Dhouibi et al. (2021).

7.3. Quantum computing

Another field in which a breakthrough is needed is the research field of quantum computing. If Europe wants to avoid what happened in 2015 (the neural MT approach from Montreal (Canada) took the MT research field by storm at WMT 2015, while until then it was mainly dominated by European universities), more research is needed on how MT, and NLP in general, can be reframed as a quantum computing problem.

Recent papers such as O'Riordan et al. (2020) and Meichanetzidis et al. (2020) outline the possibilities and challenges: "Natural language processing (NLP) problems are ubiquitous in classical computing, where they often require significant computational resources to infer sentence meanings. With the appearance of quantum computing hardware and simulators, it is worth developing methods to examine such problems on these platforms [...] NLP is often used to perform tasks such as machine translation, sentiment analysis, relationship extraction, word sense disambiguation and automatic summary generation (Cambria and White (2014)). Most traditional NLP algorithms for these problems are defined to operate over strings of words, and are commonly referred to as the 'bag of words' approach (Harris (1954)). The challenge, and thus limitation, of this approach is that the algorithms analyse sentences in a corpus based on meanings of the component words, and lack information from the grammatical rules and nuances of the language. Consequently, the qualities of results from these traditional algorithms are often unsatisfactory when the complexity of

the problem increases. On the other hand, an alternate approach called 'compositional semantics' incorporates the grammatical structure of sentences from a given language into the analysis algorithms. Compositional semantics algorithms include the information flows between words in a sentence to determine the meaning of the whole sentence (Zadrozny (1992)). One such model in this class is '(categorical) distributional compositional semantics', known as DisCoCat (Coecke et al. (2010), Zeng and Coecke (2016), Coecke (2021)), which is based on tensor product composition to give a grammatically informed algorithm that computes the meaning of sentences and phrases. This algorithm has been noted to potentially offer improvements to the quality of results, particularly for more complex sentences, in terms of memory and computational requirements. However, the main challenge in its implementation is the need for large classical computational resources. With the advent of quantum computer programming environments, both simulated and physical, a question may be whether one can exploit the available Hilbert space of such systems to carry out NLP tasks. The DisCoCat methods have a natural extension to a quantum mechanical representation, allowing for a problem to be mapped directly to this formalism (Zeng and Coecke (2016)). Using an oracle-based access pattern, one can bound the number of accesses required to create the appropriate states for use by the DisCoCat methods (Wiebe et al. (2014)). Though, this requires the use of a quantum random access memory, or qRAM (Giovannetti et al. (2008), Arunachalam et al. (2015)). Currently, qRAM remains unrealised, and expectations are that the resources necessary to realise are as challenging as a fault tolerant quantum computer (Di Matteo et al. (2020)). As such, it can be useful to examine scenarios where qRAM is not part of the architectural design of the quantum circuit. This will allow us to examine proof-of-concept methods to explore and develop use-cases later improved by its existence." O'Riordan et al. (2020).

Current work is still premature, because the hardware needed is not available yet. But it is important to note that the first theoretical steps towards reformulating MT and NLP as quantum computing problems have been made successfully. Since existing grammatical frameworks are being used to include quantum circuits in the calculation of MT and NLP problems, we believe that the European research community can pick up, provided that sufficient incentive is offered.

7.4. Context

Although MT has taken a big leap forward with the advent of neural systems, some types of translation remain very difficult. Very short sentences, for example, cause MT systems to produce inaccurate translations, because they heavily rely on context modeling: the larger the context, the better NMT systems work.

If we want MT to become pervasive for problematic text types (spreadsheets with tabular data, paragraphs, metadata fields, etc.) this problem needs addressing. For textual translation, incorporating ontological information may be of help. Continued development on multilingual lexical resources will be required for this. For multi-modal settings, extra-lingual context must be incorporated to achieve better results.

In what follows, we briefly present some MT sub-fields that can contribute to improved context modeling.

7.4.1. Unsupervised (bilingual) dictionary induction (UBDI)

For the translation of words in very small contexts (e.g., ontology labels, tabular data, metadata labels, etc.) the most popular approach nowadays consists of Unsupervised (bilingual) dictionary induction (UBDI), i. e., the derivation of multilingual vocabularies by using almost

no bilingual information. The resulting models can be used for better translating short texts, or they can be used to bootstrap MT models for low-resource languages.

In these approaches, words are organised in vector spaces, which allow similarity of words and semantic relations to be expressed, using vector operations. By linking such vector spaces in different languages, rudimentary MT systems can be created:

"A word vector space – sometimes referred to as a word embedding – associates similar words in a vocabulary with similar vectors. Learning a projection of one word vector space into another, such that similar words – across the two word embeddings – are associated with similar vectors, is useful in many contexts, with the most prominent example being the alignment of vocabularies of different languages, i. e., word translation." (Hartmann et al., 2019). This is a key step in MT of low-resource languages (Lample et al., 2017).

When constructing multilingual vector spaces, the assumption is that semantic relations are similar across languages, and hence semantically clustered vectors will look similar across embeddings. The challenge is then to find a mapping that projects as many vectors as possible with as little pre-existing information as possible (in the form of word alignments):

"Projections between word vector spaces have typically been learned from seed dictionaries. In seminal papers (Mikolov et al. (2013); Faruqui and Dyer (2014); Gouws and Søgaard (2015)), these seeds would comprise thousands of words, but Vulic and Korhonen (2016) showed that we can learn ´ reliable projections from as little as 50 words. Smith et al. (2017) and Hauer et al. (2017) subsequently showed that the seed can be replaced with just words that are identical across languages; and Artetxe et al. (2017) showed that numerals can also do the job, in some cases; both proposals removing the need for an actual dictionary. Even more recently, entirely unsupervised approaches to projecting word vector spaces onto each other have been proposed, which induce seed dictionaries in the absence of any known correspondences between words, using distribution matching techniques." (Hartmann et al., 2019).

As mentioned before, embedding projection is a good starting point for developing MT for low-resource languages, but more research into the structural organisation of embeddings is needed to better understand how structurally different languages and their respective embedding spaces can be mapped onto one another:

"Cross-lingual word embeddings aim to bridge the gap between high-resource and low-resource languages by allowing to learn multilingual word representations even without using any direct bilingual signal. The lion's share of the methods are projection based approaches that map pre-trained embeddings into a shared latent space. These methods are mostly based on the orthogonal transformation, which assumes language vector spaces to be isomorphic. However, this criterion does not necessarily hold, especially for morphologically-rich languages." (Biesialska and Costa-jussà, 2020).

Despite all progress in the field, the used methods primarily depend on the assumption that embedding spaces are homomorphic across languages. As illustrated by Biesialska and Costa-jussà (2020), further research into this topic is needed for all EU languages.

7.4.2. Document-level MT

Better context modeling is not only required to deal with very short sentences or phrases, but also to obtain more cohesive translation across larger volumes of text. Although NMT systems have successfully overcome the myopic nature of SMT (which was mainly due to the rather simplistic though effective n-gram language modeling), the state-of-the-art has not succeed yet in efficiently incorporating basic grammatical relations between sentences and paragraphs. The absence of co-reference resolution and stylistic consistency are just a few examples of features that are still missing in modern NMT systems:

"Inspite of its success, MT has been based on strong independence and locality assump-

tions, that is either translating word-by-word or phrase-by-phrase (as done by SMT) or translating sentences in isolation (as done by NMT). Text, on the contrary, does not consist of isolated, unrelated elements, but of collocated and structured group of sentences bound together by complex linguistic elements, referred to as the discourse (Jurafsky and Martin, 2009). Ignoring the inter-relations among these discourse elements results in translations which may be perfect at the sentence-level but lack crucial properties of the text hindering understanding. One way to address this issue is to exploit the underlying discourse structure of a text by utilising the information in the wider-sentential context. This is not a novel idea in itself and has been advocated by MT pioneers for decades (Bar-Hillel (1960), Sennrich (2018)), but was mostly ignored in the era of SMT due to computational efficiency and tractability concerns by the MT community. Recently, with the increase in computational power available to us and the wide-scale application of neural networks to machine translation, we are finally in a position to forego the independence constraints that have impeded the progress in MT since long." (Maruf et al., 2019).

Although some research groups have already started to look into this topic, more research into integrating discourse features is required to advance the state-of-the-art:

"Up until two years ago, there was no work in NMT that tried to incorporate any type of discourse phenomena mentioned previously, but with most sentence-based NMT systems achieving state-of-the-art performance compared to their SMT counterparts, this area of research has finally started to gain the popularity it deserves. The main difference between the research on discourse in NMT and SMT, apart from the general building blocks, is that the works in NMT rarely try to model discourse phenomena explicitly. On the contrary, they use sentences in the context directly via different modelling techniques and show how they perform on automatic evaluation while sometimes measuring the performance on specific test sets." (Maruf et al., 2019).

In addition, efforts in organising evaluation campaigns with adequate training and evaluation corpora must be increased to allow for the proper assessment of quality gains that can be achieved using document-level MT. Adapting existing corpora to fit the needs of such research may be required for this:

"Given the significant amount of work in document-level NMT in the past two years, the Fourth Conference on Machine Translation (WMT19) (Barrault et al. (2019)) and the Third Workshop on Neural Generation and Translation (WNGT 2019) (Hayashi et al. (2019)) introduced document-level translation of news and sports articles respectively, as one of the shared tasks. This opened up remarkable novelties in this domain subsuming approaches for document-level training that utilise wider document context and also document-level evaluation. To aid with this task, WMT19 produced new versions of Europarl, news-commentary, and the Rapid corpus with the document boundaries intact. They also released new versions of monolingual Newscrawl corpus containing document boundaries for English, German, and Czech. Following suit of WMT19, WNGT 2019 manually translated a portion of the RotoWire dataset2 , which contains basketball-related articles, to German. Further, they allowed any parallel and monolingual data made available by WMT19 English-German news task and the full RotoWire English dataset." (Maruf et al., 2019).

To establish the existing research as a sub-domain in its own right, it is equally important to critically research how exactly these results can be achieved. Note that this aligns with the need for explainable AI, as discussed in section 7.1.2.

An example of such study is given by Fernandes et al. (2021). The authors argue that: "while many current methods present model architectures that theoretically can use this extra context, it is often not clear how much they do actually utilize it at translation time." The paper introduces a new metric, "conditional cross-mutual information, to quantify the usage of context by these models."

7.4.3. Integrating visual features for MT

In the previous sections, we have discussed required breakthroughs for single-mode translation for written text. With the exception of very short sentences and phrases, for most of such text types, a sufficiently large context is provided for NMT systems to work well. However, the majority of human language is produced outside of written texts. In such settings, extra-lingual cues are often required to decode a message adequately and to translate it correctly. To enable better modeling of multi-modal environments, we not only need research into how modalities can enrich one another, but also in how training and test sets can be constructed to achieve better modeling. Sanabria et al. (2018) expresses this problems as follows: "Multimodal sensory integration is an important aspect of human concept representation, language processing and reasoning (Barsalou et al. (2003)). From a computational perspective, major breakthroughs in natural language processing (NLP), computer vision (CV), and automatic speech recognition (ASR) have resulted in improvements in a wide range of multimodal tasks, including visual question-answering (Antol et al. (2015)), multimodal MT (Specia et al. (2016)), visual dialogue (Das et al. (2017)), and grounded ASR (Palaskar et al. (2018)). Despite these advances, state-of-the-art computational models are nowhere near integrating multiple modalities as effectively as humans. This can be partially attributed to a lack of resources that are pervasively multimodal: existing datasets are typically focused on a single task, e. g., images and text for image captioning (Chen et al. (2015)), images and text for visual-question answering (Antol et al. (2015)), or speech and text for ASR (Godfrey et al. (1992)). These datasets play a crucial role in the development of their fields, but their single-task nature limits the collective ability to develop general purpose artificial intelligence."

In line with our remark in section 7.4.2, it is of critical importance to gain a better understanding of how visual features may improve MT and to develop analysis methods that unambiguously demonstrate improvements over baselines. Caglayan et al. (2019), for example, shows through an ablation study that visual context does contribute to better translation:

"Current work on multimodal machine translation (MMT) has suggested that the visual modality is either unnecessary or only marginally beneficial. We posit that this is a consequence of the very simple, short and repetitive sentences used in the only available dataset for the task (Multi30K), rendering the source text sufficient as context. In the general case, however, we believe that it is possible to combine visual and textual information in order to ground translations. In this paper we probe the contribution of the visual modality to state-of-the-art MMT models by conducting a systematic analysis where we partially deprive the models from source-side textual context. Our results show that under limited textual context, models are capable of leveraging the visual input to generate better translations. This contradicts the current belief that MMT models disregard the visual modality because of either the quality of the image features or the way they are integrated into the model."

Note that this study again emphasises, as previously discussed in section 7.4.2, that appropriate data sets for training are required.

7.4.4. Integrating audio features for MT

Following the same logic as for visual features, research groups have begun to also add audio context to their NMT systems. "Evidence from human learning suggests that additional modalities can provide disambiguating signals crucial for many language tasks" (Specia et al., 2020), because "embedding vectors are sensitive to the meaning of words and allow semantically similar words to be near each other in the vector spaces and share their statistical power. Unfortunately, the model often maps such similar words too closely, which complicates distinguishing them. Consequently, NMT systems often mistranslate words that seem natural in the context but do not reflect the content of the source sentence. Incorporating auxiliary information usually enhances the discriminability." (Kano et al., 2019).

The term “acoustic word embedding” is used to denote “fixed-dimensional representations of variable-length speech segments.” (Kamper et al., 2021). In recent work, (Deena et al. (2017)) “auxiliary features derived from accompanying audio, are investigated for NMT and are compared and combined with text-derived features. These acoustic embeddings can help resolve ambiguity in the translation, thus improving the output. The following features are experimented with: Latent Dirichlet Allocation (LDA) topic vectors and GMM subspace i-vectors derived from audio. These are contrasted against: skip-gram/Word2Vec features and LDA features derived from text. The results are encouraging and show that acoustic information does help with NMT, leading to an overall 3.3% relative improvement in BLEU scores.”

For low-resource languages, experiments with multilingual transfer have been conducted. (Kamper et al. (2021)), for example, investigated “zero-resource settings where unlabelled speech is the only available resource”. The authors apply “a single supervised embedding model on labelled data from multiple well-resourced languages and then apply it to unseen zero-resource languages” using “three multilingual recurrent neural network (RNN) models: a classifier trained on the joint vocabularies of all training languages; a Siamese RNN trained to discriminate between same and different words from multiple languages; and a correspondence autoencoder (CAE) RNN trained to reconstruct word pairs. In a word discrimination task on six target languages, all of these models outperform state-of-the-art unsupervised models trained on the zero-resource languages themselves, giving relative improvements of more than 30% in average precision. When using only a few training languages, the multilingual CAE performs better, but with more training languages the other multilingual models perform similarly. Using more training languages is generally beneficial, but improvements are marginal on some languages.”(Kamper et al. (2021))

Similarly to visual features, introspective research is required into “the composition of features in different settings in order to better understand the type of complementary information they bring and how these can be leveraged effectively in NMT systems” (Deena et al. (2017)) and the development of suitable data sets is required. (Specia et al. (2020)) describe, for example, “the How2 data set, a large, open-domain collection of videos with transcriptions and their translations [... and ...] show how this single data set can be used to develop systems for a variety of language tasks and present a number of models meant as starting points [...] This corpus brings together English audio, English transcripts, Portuguese transcripts, videos, and summaries, along with meta-data such as topic of the video. This makes the How2 data set a good resource for research at the intersection of vision, language and speech.

7.5. End-to-end MT

7.5.1. End-to-end speech translation

Recently, the need for research into end-to-end systems has been widely recognised. For example Inaguma et al. (2020) discuss the advantages end-to-end systems have over cascaded systems in terms of inference speed and error reduction: “Speech translation (ST), where converting speech signals in a language to text in another language, is a key technique to break the language barrier for human communication. Traditional ST systems involve cascading automatic speech recognition (ASR), text normalization (e.g., punctuation insertion, case restoration), and machine translation (MT) modules; we call this CascadeST (Ney (1999); Casacuberta et al. (2008); Kumar et al. (2014)). Recently, sequence-to-sequence (S2S) models have become the method of choice in implementing both the ASR and MT modules (c.f. Chan et al. (2016); Bahdanau et al. (2014b)). This convergence of models has opened up the possibility of designing end-to-end speech translation (E2E-ST) systems, where a single S2S directly maps speech in a source language to its translation in the target language (Bérard et al. (2016)

; Weiss et al. (2017)). E2E-ST has several advantages over the cascaded approach: (1) a single E2E-ST model can reduce latency at inference time, which is useful for time-critical use cases like simultaneous interpretation. (2) A single model enables back-propagation training in an end-to-end fashion, which mitigates the risk of error propagation by cascaded modules. (3) In certain use cases such as endangered language documentation (Bird et al. (2014)), source speech and target text translation (without the intermediate source text transcript) might be easier to obtain, necessitating the adoption of E2E-ST models (Anastasopoulos and Chiang (2018)). Nevertheless, the verdict is still out on the comparison of translation quality between E2E-ST and Cascade-ST. Some empirical results favor E2E (Weiss et al. (2017)) while others favor Cascade (Niehues et al. (2019)); the conclusion also depends on the nuances of the training data condition (Sperber et al. (2019)).”

Since both approaches have their merit, it is probably too early to fully commit to either of the paradigms. For this reason Inaguma et al. (2020) propose a unified framework. They “believe the time is ripe to develop a unified toolkit that facilitates research in both E2E and cascaded approaches.” and present ESPnetST, a toolkit that implements many of the recent models for E2E-ST, as well as the ASR and MT modules for Cascade-ST. the goal is “to provide a toolkit where researchers can easily incorporate and test new ideas under different approaches. Recent research suggests that pre-training, multi-task learning, and transfer learning are important techniques for achieving improved results for E2EST (Bérard et al. (2018); Anastasopoulos and Chiang (2018); Bansal et al. (2018); Inaguma et al. (2019)). Thus, a unified toolkit that enables researchers to seamlessly mix-and-match different ASR and MT models in training both E2E-ST and Cascade-ST systems would facilitate research in the field. There exist many excellent toolkits that support both ASR and MT tasks [...]. However, it is not always straightforward to use them for E2E-ST and Cascade-ST, due to incompatible training/inference pipelines in different modules or lack of detailed preprocessing/training scripts.”

While such framework will undoubtedly contribute to advancing the state-of-the-art in developing end-to-end MT systems for spoken language in an experimental way, it is also important to keep emphasising the importance of introspective research that looks into the root causes of MT performance issues. Sperber and Paulik (2020) start with providing a “categorization of ST research into well-defined terms for the particular challenges, requirements, and techniques that are being addressed or used. This multidimensional categorization suggests a modeling space with many intermediate points, rather than a dichotomy of cascaded vs. end-to-end models, and reveals a number of trade-offs between different modeling choices. This implies that additional work to more explicitly analyze the interactions between these trade-offs, along with further model explorations, can help to determine more favorable points in the modeling space, and ultimately the most favorable model for a specific ST application.”

“By contrasting the extreme points of loosely coupled cascades vs. purely end-to-end trained direct models,” Sperber and Paulik (2020) “identify foundational challenges: erroneous early decisions, mismatch between spokenstyle ASR outputs and written-style MT inputs, and loss of speech information (e. g., prosody) on the one hand, and data scarcity on the other hand.” The authors “then show that to improve data efficiency, most end-to-end models employ techniques that re-introduce issues generally attributed to cascaded ST.”

Based on an in-depth analysis of various architectures, Sperber and Paulik (2020) suggest pathways for further research. They “conjecture that the apparent trade-off between data efficiency and modeling power may explain the mixed success in outperforming the loosely coupled cascade. In order to make progress in this regard, the involved issues (early decisions, mismatched source-language, information loss, data efficiency) need to be precisely analyzed [...], and more model variants [...] should be explored. As a possible starting point one may aim to extend, rather than alter, traditional models, e. g., applying end-to-end training as a fine-tuning step, employing a direct model for rescoring, or adding a triangle con-

nection to a loosely coupled cascade." The authors "further suggest that more principled solutions to the different application-specific requirements [...] should be attempted. Perhaps it is possible to get rid of segmentation as a separate step in batch delivery mode, or perhaps text as output medium can be used to visualize repairs more effectively. Several of the application-specific requirements demand user studies and will not be sufficiently solved by relying on automatic metrics only."

7.6. Explainability

The problem of explainable MT is described by Stahlberg et al. (2018): "Neural machine translation (NMT) models (Sutskever et al. (2014); Bahdanau et al. (2014a); Gehring et al. (2017); Vaswani et al. (2017)) are remarkably effective in modelling the distribution over target sentences conditioned on the source sentence, and yield superior translation performance compared to traditional statistical machine translation (SMT) on many language pairs. However, it is often difficult to extract a comprehensible explanation for the predictions of these models as information in the network is represented by real-valued vectors or matrices (Ding et al. (2017)).

At the very core of the problem is the internal representations that are being used by NMT systems: "A trend in NMT has been to abandon [traditional token-based pipelines] in favour of an end-to-end procedure Lee et al. (2016). Thus some of the best-performing NMT systems convert character strings of the source language to character strings of the target language. They don't assume even the most rudimentary linguistic abstractions, such as words. From a theoretical AI point of view, this makes some sense: end-to-end systems may be better models of the intuitive behaviour of humans, under the assumption that humans' knowledge of language is implicit and not based on linguistic concepts. From the engineering point of view, however, end-to-end translation is problematic:

- it can make mistakes that could be easily avoided by well-established linguistic knowledge;
- its functioning is hard to understand, predict, and improve;
- it leaves no trace (intelligible to humans) of how the translations are obtained.

By not using established linguistic knowledge, an end-to-end black box system is in conflict with the scientific ideal of assuming as little as possible and proving as much as possible. In its simplest form, this ideal says: don't guess if you know Tapanainen and Voutilainen (1994). It is a virtue of the traditional pipeline that it builds on known things and minimizes guessing. From this pipeline, it is moreover possible to extract explanations of decisions made at different stages. These explanations can help engineers to improve the program and users to assess the reliability of each translation." Ranta (2017).

While alternative MT approaches may have weaker performances, they had the advantage of being "transparent", i. e., users and researchers could diagnose translation problems by looking at the inner workings of MT systems: "In contrast, the translation process in SMT is 'transparent' as it can identify the source word which caused a target word through word alignment. Most NMT models do not use the concept of word alignment." Stahlberg et al. (2018). As a result, "Explainable and interpretable machine learning is attracting more and more attention in the research community (Ribeiro et al., 2016; Doshi-Velez and Kim, 2017), particularly in the context of natural language processing (Karpathy et al., 2015; Li et al., 2016; Alvarez-Melis and Jaakkola, 2017; Ding et al., 2017; Feng et al., 2018). These approaches aim to explain (the predictions of) an existing model." Stahlberg et al. (2018).

7.7. Training data

From previous sections, it has become clear that having the appropriate data sets is critical for the further development of NMT systems, not only for training, but also for analysing the contribution of features to the overall translation quality of multi-modal systems.

We discern two important breakthroughs that must be achieved:

- Creation of new data sets, re-iteration over existing data sets
- Support from a policy for public data re-use

7.7.1. Creation of new data sets, re-iteration over existing data sets

(Specia et al. (2020)) has demonstrated how a multi-modal corpus can contribute to the development of multiple NLP tasks, and already marks a good starting point for further data collection efforts. However, the number of languages available in the How2 data set is very limited (only English and Portuguese are available), and the extracted number of words is well below that of most languages in the Europarl (Koehn et al. (2005)) data set (roughly at 50%). In addition, translations have been crowdsourced, so it can be assumed that the quality is not on par with the professionally translated Europarl corpus. It is most interesting though, to learn that, given the original English transcriptions that accompany the video material, it took less than US$10,000 to complete the crowdsourcing campaign for one language pair. The proposed methodology (which provides mechanisms for filtering out low-quality contributions), perhaps combined with a form of gamification (see for example Eryiğit et al. (2021)), can be applied to develop new language pairs.

Ideally, new data annotation efforts should build further upon existing work. For document-level NMT this can be done with limited effort, as demonstrated in the WMT19 campaign (Barrault et al. (2019)). For video and audio content, it will most definitely require more work, but with the existing NLP technology it is not unthinkable that Europarl sessions can be semi-automatically linked with related video and audio content to create an annotated corpus that can be used for both building new NMT systems and analysing the contribution of features to translation quality.

7.7.2. Support from a policy for public data re-use

As already outlined in 7.1.3, there are policies in place that encourage the release of public data for re-use by academia and industry (see, for example, COM(2018) 234 final). Unfortunately, most of these policies seem to exclusively relate to machine-readable data that is strictly governed by data standards, i. e., sharing is easily institutionalised because the defining standards prevent any possible conflict with the GDPR. Conversely, any human-readable data seems to be governed by the GDPR, because one can never predict whether sensitive personal data is involved. As a result, the LT community is burdened with complex procedures (requesting permission, legal screening, anonymisation, ...) for the consumption and production of data.

Such procedures could be easily avoided if public organisations (government administrations, public broadcasting services, ...) were encouraged to publish their data using a very simple data specification. To begin with, such specification could signal potential consumers that a document does not contain any sensitive personal data, and thus no extra efforts are required for consumption. In its simplest form, such specification would already be of great help for the LT community. In a more elaborate form, such specification could also be beneficial for the public organisations themselves, because it would allow them to make their data more structured and accessible, so it can be re-used for other purposes.

The irony of it all is that such (unofficial) specification is already widely used: Search Engine Optimisation (SEO) is often implemented on web sites to make it easier for search engines to index relevant information. Moreover, a whole industry of SEO engineering has emerged with the sole purpose of constantly optimising content annotations for the benefit of the technology giants.

8. Machine Translation: Main Technology Visions and Development Goals

8.1. Models and systems

8.1.1. Model size

The strategy of **building huge MT models** by comprising all available data coming from many different domains (and also languages in current multilingual systems) should be complemented by **developing smaller models**, too. These small(er) models should be trained using the largest possible set of available information, helping under-resourced languages and domains by appealing to knowledge from higher-resourced ones. One of the current problems is that if this results in a single huge model, most practitioners cannot run the model owing to hardware constraints. Accordingly, smaller models adapted to particular language pairs and genres/domains should be made available. Following the foreboding claims by Strubell et al. (2019b) of the negative effect on climate change by the CO_2 produced by our large models, Jooste et al. (2022) show for a specific MT service provider that 'student' models distilled down from larger 'teacher' models (Kim and Rush, 2016) can reduce emissions, as well as cost purely in monetary terms, by almost 50%, with little decrease in quality. Of course, such techniques benefit not only MT, but NLP as a whole: they could be used individually for specific language pairs, domains and applications, as well as jointly if translation in a general domain is required. This would have several benefits: such models would be easy to integrate and use on any device (e. g., mobile phones), provide high-quality translations for all domains and languages, and also be greener by requiring fewer resources. If more models were to be used on one device (e. g., if the user wants to be able to translate restaurant menus and information about places to visit), automatic selection of the appropriate domain according to the user's request should be available. These aspects are important since improvements in the quality of the translations produced coupled with improvements in the accessibility of the tools through the development of mobile applications are now allowing users to translate from text, speech and images, which has led to a popularisation of the systems among the public at large, not just in their respective communities.

8.1.2. Availability

Note too that future **publicly available MT systems** should not depend on large companies, especially those which are not European. The risk is that what is freely available now (e. g., Google Translate,[41] Bing Translator,[42] etc.) could (easily) be taken away if those companies – none of them MT companies *per se*, note – find a way to increase revenue in other directions, so that they deprecate their MT offerings, as has happened with other services provided by these large corporations.

[41] https://translate.google.com
[42] https://www.bing.com/translator

8.1.3. Bias

Another challenge of the current systems is represented by various **biases** in the models, such as gender, racial and ethnic bias (Vanmassenhove and Way, 2018; Vanmassenhove et al., 2019). These biases replicate regrettable patterns of socio-economic domination that are conveyed through language, since these biases are present in the training data and are then amplified by models which tend to choose more frequent patterns and discard rare ones. In the future, ethical and fair MT should not further propagate notions of inequality, but rather foster an inclusive society based on acceptance and respect: in future models, those biases should be removed altogether, to ensure that the language produced by such systems does not reinforce and propagate inequality and exclusion. One way to achieve this is the examination of training data, identifying biased parts or gaps, and enriching the data by providing alternatives, or by replacing them altogether. Modifying models could reduce biases, too, for example by introducing weights for probabilities of words related to bias.

8.1.4. Context

Note too that more and more NMT systems are being developed which go beyond the single sentence level (Huo et al., 2020; Lopes et al., 2020), using a variety of different approaches: taking source language context, taking target level context, or both. Another interesting avenue being pursued is that different context spans have been investigated, ranging from a single preceding sentence to the entire "document". Often, context-aware systems and evaluation are referred to as "document-level", although it is still not clear what the "document" exactly is. While this might be straightforward for news articles and user reviews, the situation is different for literary texts or movie subtitles, to name but two. Some sentences are translated well regardless of the context in which they appear, but future systems should be able to identify which sentences benefit from the availability of context, and to be able to find that context in the document. This is far from trivial because context-relevant information can be found in different places, and is encoded differently, more or less excplicitly, across languages: sometimes it is indeed in the immediately preceding sentence, but often it is further back, or even in some cases in sentences which follow on from the sentence currently being translated. Sometimes it is even necessary to get access to data going beyond the given text, such as the topic of the text, the gender of the writer/speaker, or even world knowledge more generally.

8.1.5. Multimodal models

The need to move towards context-aware methodologies in MT is indisputable, both for building systems and for evaluation. A new definition of how context should be addressed is essential, including understanding the context-related issues in different languages and domains, and the context span necessary to solve those issues, including external information (meta data). This external information can go beyond text data and include images, videos, tables, etc. by developing **multimodal MT systems** (Yao and Wan, 2020). Such systems currently include image information to help in the translation of image captions. Future systems should combine different sources of information going beyond this, such as an image of a product should help to disambiguate words in the description or review of this product, etc.

Given that information (e. g., on the Internet and social media in particular) is increasingly disseminated by combining text and images, as well as through media such as videos and podcasts, there is an urgent need to improve MT systems that can support the meaningful integration of the written and spoken word with images. In particular, the visual element can help to disambiguate meaning and provide much-needed context for the translation.

Multimodal models should also include sign language translation, which is currently relying mainly on computer vision methods. Sign language translation should use the models based both on images and on natural language.

8.2. Data

8.2.1. Training data

In addition to test sets, which are crucial for assessing the systems, **training data** crucial to building models should receive more attention, too. Currently, the majority of MT systems are trained on large amounts of data covering only a small amount of languages, language pairs and domains. While the progress of MT is mainly measured on high-resource conditions, the majority of domains and languages, including many of those spoken in Europe, are **under- or low-resourced**. Future systems should be able to cover all European languages as well as language pairs (not always including English or some other higher-resourced language), and be trained on many different domains and genres. For this to work for all, and not only for big companies and leading research teams to be able to benefit from large amounts of data, the availability and quality of training data should be increased. Publicly available multilingual data should include a greater diversity of domains and languages so that building high-quality MT systems becomes an option for all.

Multilingual models and zero-shot MT, unsupervised MT, synthetic data, transfer learning. Still, if there is not enough data for a language (pair), and even monolingual data can be scarce for some language, all these methods will not reach the final goal. Novel methods and research breakthrough are needed in this direction.

8.2.2. Test sets

Currently, the **test sets** predominantly used in a large number of research publications are those coming from the WMT[43] shared task. The researchers are testing their systems on these texts reporting improvements of automatic scores (mainly BLEU). However, some of the human translations in these test sets used as references for automatic scores are found to be of poor quality (Toral et al., 2018). Furthermore, some of them are found to be post-edited MT outputs. Therefore, reporting improvements by a small amount of BLEU points on those test sets does not necessarily mean that the new system is improving; it might even mean that the system is deteriorating. The shared task organisers cannot be blamed for this situation, as they do the best that they can with the limited budgets that they have. Still, these human translations should be thoroughly examined in order to discard the inappropriate ones and keep only the good ones for the long-term testing of systems with these validated data sets. Note that in the light of comparison between MT outputs and human translations carried out frequently in recent years and often claiming “human parity”, the quality of human translations has to be high.

Expanding the set of test sets In addition, **new test sets** coming from *different genres and domains* need to be much more widely used. A vast amount of systems are currently tested only on a limited set of domains, news being the predominant one in the WMT shared task. While other domains are also emerging in shared tasks (e. g., biomedical at WMT, and formal speech in form of TED talks at IWSLT), many other genres and domains are as yet hardly covered by current research, such as user-generated content, which is also not a homogeneous genre (consisting of less noisy user reviews coming from distinct topics, noisy

[43] http://www.statmt.org/wmt21/

social media posts, conversations on forums, WhatsApp exchanges, etc.), but which probably has the greatest potential for future growth and poses the most complex challenges. In the long-term, then, we strongly contend that MT systems should be tested on a large number of different domains and genres, and for an ever-increasing range of languages in order to help facilitate DLE.

Challenge test sets Furthermore, the rise of NMT and its increasing quality has led to more and more **challenge test sets (or test suites)** being developed in order to better understand the particular strengths and weaknesses of systems. These specified test sets enable better understanding of certain (linguistic) aspects which cannot be properly assessed in standard "natural" test sets. Still, the development and creation of such test sets necessitates a large extent of human expertise, time and effort. In the future, these test sets should be easy and fast to create for any language pair, given the desired aspect to be tested (Arnold et al., 1993).

8.3. Evaluation

As for the **evaluation** process, automatic evaluation metrics still remain invaluable tools for rapid development and comparison of MT systems. They have been developed and improved constantly, with more and more metrics coming up each year. However, a number of challenges remain. One thing is that the community still relies to a large extent on one of the first automatic metrics, BLEU, and there is a noticeable reluctance to abandon this measure despite a large body of research pointing out its drawbacks Mathur et al. (2020); Kocmi et al. (2021). Future systems should be evaluated by new automatic metrics which represent better approximations of human judgments and also ideally abandon the dependence on human reference translations, which is a serious limitation. While the quality of human translations can be revised and controlled in the future as mentioned above, the problem of scores providing information only on how close the system output is to just a single possible correct translation among many does not essentially reflect any actual aspect of translation quality in the real world. Future metrics should be designed in a flexible manner so as to use the original text and MT output to provide information about the desired quality aspects for the specific task at hand.

8.3.1. Manual evaluation

Manual evaluation of translation quality, despite its disadvantages (time- and resource-intensive, as well as being subjective), remains the gold standard, both for evaluating MT systems as well as for developing suitable automatic metrics. That being said, the design of experiments and the standard method of reporting the results is far from perfect. Different papers use the same quality criterion name with different definitions, or the same definition with different names. Furthermore, many papers do not use any particular criterion, asking the evaluators only to assess "how good" the output is. As Way (2018) notes, "If there ever truly was a single notion of quality as regards translation – namely 'perfect' human translation – then this needs to be abandoned forthwith; the range of situations in which MT is being deployed nowadays includes many where there simply is no place for human intervention, either in terms of speed, or cost, or both". In Way (2013), he suggests that "Each of the services facilitated by MT will have its own definitions of quality, dependent on the client's content and business requirements. Quality will be able to be assessed by end-users or buyers, instead of in-country reviewers."

Quality criteria We agree that any idea of a single standard general unspecified notion of quality should be abandoned, and factors like the context in which MT is to be used together

with appropriate quality aspects should be considered. Quality needs to be measured according to an agreed specification outlining the specific criteria the MT output needs to meet, in order to have met the level of required quality for the task at hand. These criteria might include adequacy/accuracy, readability/comprehension, appropriate register, correct terminology, consistency (with a reference point), acceptability to the end user, adequately fulfilling a function (e. g., preserving humour or sentiment, providing instructions to correctly complete a guided task, etc.). However, such elements are missing from current MT quality assessment methods as conducting all these required steps is non-trivial. It is much easier to get a bilingual member of the crowd to say whether the translation is "good", "adequate", or "poor". Consequently, automatic metrics should be created with criteria like these designed in from the outset, and not only to provide a general unspecified score which is meaningless to most people.

Furthermore, recent research has found that readers tend to fully trust fluent translations as well as comprehensible translations even if they contain severe adequacy errors which change the actual content and deliver completely different information (Martindale and Carpuat, 2018; Popović, 2020; Martindale et al., 2021). Therefore, future automatic metrics should provide confidence indicators for translations in order to inform users about the level of trust they should have in the MT output they are reading.

Context Recent research showed that both manual and automatic evaluation should not be carried out on isolated sentences/segments, but rather taking into account the broader **context** of the document at hand in order to be more reliable (Läubli et al., 2018; Castilho, 2021). Context-aware evaluation of MT is of interest to the community as it enables the assessment of supra-sentential context which, in turn, provides more meaningful insights into the actual quality of the MT output. Therefore, context should always be used in future manual evaluations as well as in future automatic metrics.

8.3.2. Automatic evaluation

Currently, more and more **automatic metrics** based on neural networks and/or word representations are emerging which show better correlations with human judgments. However, these metrics still require training data which, similarly to the training data for building MT systems, are available only for a limited number of language pairs and domains. Future automatic metrics should be equally valid without such constraints.

Furthermore, as mentioned in 8.3.1, increased attention should be paid to the human judgments used for tailoring the automatic metrics, as well as to manual evaluation in general.

8.4. Applications

8.4.1. Spoken-language translation

Allowing users to interact naturally with machines via speech has the potential to greatly transform, enhance and empower work, leisure and social experiences. The increasing quality of MT and the expanding preference (especially among younger users) for voice-based interaction with devices points to more and more applications for **speech-to-text** and **speech-to-speech translation**, e. g., a user should be able to go into a doctor's surgery in a foreign country and use reliable spoken-language translation in a medical setting, or book a hotel room in a country where they do not speak the language. More generally, the number of queries being conducted via voice as opposed to the more traditional method of typing in keywords into a search engine is increasing all the time. This means, of course, not only that spoken language input should become more and more a topic of close attention, but

also that more data of exactly the right type is becoming available. By 2030, it is likely that the Automatic Speech Recognition-MT-Speech Synthesis pipeline will have been replaced by more direct approaches which model spoken language translation as an end-to-end process (Di Gangi et al., 2019), but clearly more work needs to be done in this regard.

8.4.2. Sign language translation

Sign language translation should be widely available for many domains (medical, touristic, etc.) in order to break down language barriers for deaf and hear impaired users and enable them to access all information. As previously mentioned, sign language translation should include language features in addition to image features. In addition, it should not only be translated from and into text but also from and into speech, by including speech features into its multi-modal models.

8.4.3. Language learning

Another aspect underpinning the expanding popularisation of MT is the fact that **language learners** are using freely available web-based MT systems to carry out coursework or to communicate in a second language. Language teachers argue that the use of MT for language learning purposes can weaken learners' autonomy in a second language. However, recent research (Resende and Way, 2021) has shown that interaction with the MT output can be beneficial to language learners as the MT output facilitates the cognitive processing and acquisition of challenging structures in a second language. Considering this scenario, more research on the effects of MT systems on their end-users is necessary for MT developers and researchers to understand how the overall quality of the system can be improved to be compatible with human quality, thus providing more engaging texts in the target language. Again, systems for this purpose should also rely on smaller models and be independent of large multinational companies. Furthermore, research on the effects of MT on language learning and processing is necessary to inform language teachers as to how they can integrate these systems into their teaching activities.

8.4.4. Multilingual NLP tasks

MT is more and more being used for **expanding other NLP tasks** (e. g., text classification, topic modelling, sentiment analysis) to multiple languages. Usually, full translation is carried out and then the labels for the original source language together with the translations are used for training classifiers in the new target language. However, for such tasks, where the translated text is not used directly, quality criteria might be rather different, and full translation might not even be necessary. Extracting different representations from various layers could be even better suited for certain tasks. This option should be made easily available in future MT systems, with all representations needed for the task at hand provided in an easily reusable format.

9. Machine Translation: Towards Deep Natural Language Understanding

Texts serve a communicative purpose. They are not only written to transmit factual knowledge. Rather, they are also meant to initiate (or stop) actions, evoke emotions, or simply entertain. They only succeed in doing so when they hit a sweet spot that amongst others is

determined by societal, situational, and personal context. Translation is not only the business of analysing a source text sentence by sentence. Rather, the author and the intended audience as well as the purpose of communication – the interactional, communicative context – need to be considered as well.[44] Accordingly, MT needs to be assessed with regard to questions such as "Do texts generated by MT fulfill the intended communicative purposes?", and ultimately "Does the intended communication succeed?".

Applying this purpose- and communication-oriented view on MT facilitates a discussion of the question "Does MT need Natural Language Understanding?" or even "Does MT need Deep Natural Language Understanding?", since it helps to put the prevailing MT-related metrics in perspective. These metrics – referred to by acronyms like BLEU, hLEPOR, BERTScore, and COMET – are not related to purpose and communication aspects. They are not even about verifiable qualities of translations like accuracy, or adherence to a translation-related specification (e.g., demand to transcreate humour, or to subjective criteria such as "artistry" or "elegance").[45] The existing MT-related evaluation metrics measure nothing but similarity to a reference translation unit/pair that does already exist. Accordingly, claims related to MT reaching parity with human translations are misleading since the metrics to measure this via reference translation data are too limited to answer the question "Does the intended communication succeed/fulfill its purpose?" or similar intricated questions related to reader impression, and style.

With a view on communication success, it becomes obvious that MT – core technology, evaluation methodologies, metrics and data for training and evaluation – needs NLP that goes beyond traditional capabilities such as detection of terms / keywords / labels, entities, relations, and sentiments. These capabilities – amongst others referred to as Deep Natural Language Understanding will be

- aware of context
- able to consider annotations/metadata

Context- and annotation-awareness will allow MT to generate texts that

- are faithful to the intended communication (input view)
- take translation purpose/specifications/requirements into account (sender view)
- show empathy with the reader/listener (output/consumer view)

Only MT with Deep Natural Language Understanding will for example be able to efficiently support a human-to-human or human-to-machine conversation (chatbot/digital assistant dialogue) that exhibits qualities like the following:[46]

1. *Contextualized* – sensitive to
    - input from all modalities (e.g., tailor real-time speech-to-speech MT on visuals or slides that a presenter uses, or adapt MT according to prosodic cues from speech input)[47]

[44] This focus on the *purpose* of a translation activity is a key aspect of the Skopos theory and functional translation perspective (developed and refined by Reiss, Vermeer, Kussmaul, Nord, and others) which is one of the prevalent approaches taught in human translation studies.

[45] See http://www.qt21.eu/mqm-definition/definition-2015-12-30.html

[46] See suggestions related to the "Responsive MT" approach sketched by Arle Lommel (Common Sense Advisory) in https://csa-research.com/Blogs-Events/Blog/responsive-MT-Test.

[47] Expanding MT solutions beyond language resources by integrating multimodal capabilities (see https://amtaweb.org/wp-content/uploads/2021/10/Specia-keynote-SiMT_MTSummit_LS.pdf) would even allow to increase traditional quality and adequacy of MT output.

- ingested annotations/metadata (e.g., domain-, industry- or company-specific terminology)
- surrounding content (e.g., previous utterance in a human-to-machine conversation)

2. *Adaptive* – taking into account
   - various sources of feedback (e.g., explicit via prompts from chatbots such as "Did this answer your question?")
   - language evolution (e.g., reflecting introduction of new vocabulary as experienced during Covid)
3. *Personalized* – tailored to readers'/listeners'
   - general cognitive capabilities (e.g., difficulties to interpret long words or complex sentences[48])
   - linguistic skills and preferences (e.g., colloquial style)
   - interest (e.g., get a short summary of a text)
   - situational consumption environment (e.g., receiving spoken output, or even an image or video
   - cultural background, and actual location (e.g., providing a French translation for taxes relevant in France to provide locale-relevant content in a globalised world)
   - Inclusive (e.g., gender aware[49])
   - Accessible (e.g., considering hearing or vision impaired content consumers)
4. *Knowledge*-rich – considering
   - external information (e.g., available for example based on Linked Data principles, and persisted in Knowledge Graphs)
   - possible constraints, and inferences (e.g., encoded in Shapes Constraint Language (SHACL)[50] and drawn via a reasoner processing the Web Ontology Language (OWL)[51]

The following ingredients currently seem to emerge as important elements for a next-generation MT (based on Deep Natural Language Understanding):

- Existing standards related to annotations,[52] texts (e.g., XLIFF for bilingual texts), languages, and concepts will be the backbone required for Deep Natural Language Understanding.
- The FAIR data principles[53] will guarantee maximum leverage for data and help to adhere to data management recommendations from the European Commission. They also are valuable or even indispensable in the context of GDPR-compliance, and ethical guidance – also denominated "responsible MT"[54].

[48] See e. g., https://www.easy-plain-accessible.com for language variants addressing this
[49] See considerations on "GenderFairMT" (in German): https://www.universitas.org/genderfaire-sprache-und-sprachtechnologie-genderfairmt/
[50] See https://www.w3.org/TR/shacl/
[51] See https://www.w3.org/TR/owl2-primer/
[52] Questions on how to integrate which type of meta-data into bi-text corpora and how these could best be leveraged by MT models still need to be explored and answered, see e. g. Alan Melby's statements in https://multilingual.com/issues/november-december-2021/data-of-course-mt-useful-or-risky-translators-here-to-stay/. Inspirations and suggestions can be found e.g., in Lieske (2020)
[53] See https://www.go-fair.org/fair-principles/
[54] Arle Lommel (Common Sense Advisory, CSA)

- Experts like translators, domain specialists, modelers, data scientists will remain important contributors to create, revise, and assess the data and artefacts that constitute the foundations of modern MT.
- Open, standardised, flexible and robust technologies for data management, data production, data cleaning, data labelling, data classification, data curation, data analysis and data quality checking.
- Large, multilingual translation models[55] that are "safe to use" and can be adapted (e.g., using transfer learning) easily for resource sparse computing environments, and to specific tasks and domains will allow advances for low-resource languages.

10. Summary and Conclusions

This deep dive has provided a condensed summary of the current state of the art in the field of MT and has suggested recommendations and directions for expected and desirable developments going towards 2030, especially to ensure that MT contributes to achieving DLE for all the languages of Europe.

From the beginning the main goal of MT has been to provide high-quality, robust translation between any language pair. Today translation technologies are widely used the by general public, public sector and government agencies, SMEs, LSPs and many other industries where multilingual content is indispensable. The use of translation technology will definitely continue growing, covering new application areas (e.g., Internet of Things, smart homes and other smart devices), markets, supporting Europe's digital single market and language equality. When looking forward to 2030, we expect the movement towards Deep Natural Language Understanding enabling efficient and real-time translation to support human-to-human or human-to-machine communication.

Despite the widespread celebration of multilingualism in the EU, there is no common policy addressing language barriers. There is also a gap in publicly available MT services which cater specifically to the needs of people in Europe. Users around the world avail of free-of-charge MT services provided by Google[56], Microsoft[57], Baidu, etc. The risk is that what is freely available now could (easily) be taken away if those companies find a way to increase revenue in other directions, as has happened with other services provided by these large corporations. The absence of a clear roadmap and support for LT at European level translates into an incohesive, fragmented European market with disparate language support for the language communities of Europe. The future **publicly available MT systems** should not depend on large companies, especially those which **are** not **European**.

With the help of neural networks, MT has recently improved significantly in its quality, consistency and productivity. However, in many cases the focus of new technologies is still on big, fully-resourced languages, in particular English, thus limiting diversity and reinforcing already-existing disparities. At the same time the neural network techniques have opened the path to developing a universal translation engine aiming to translate between any language pair with help of a single model. The application of neural networks to MT allows also to forego the independence constraints and move towards context-aware methodologies in MT. A novel approach attracting the attention of many researchers is unsupervised MT, where (every time less) monolingual data suffices to build a working system. While much work remains to be done in this area, together with universal MT, it emerges as one of the key pillars to drive language equality.

[55] See e.g., https://ai.facebook.com/research/publications/emerging-cross-lingual-structure-in-pretrained-language-models/
[56] https://translate.google.com
[57] https://www.bing.com/translator

Another challenge of the current systems are various **biases** in the models, such as gender, racial and ethnic bias. In the future, ethical and fair MT should not further propagate notions of inequality or exclusion, but rather foster an inclusive society.

Explainable and interpretable machine learning is attracting more and more attention in the research community. A fundamental breakthrough is needed in the understanding of how current MT algorithms work.

Another field in which a breakthrough is needed is quantum computing. Finally, first theoretical steps towards reformulating MT and NLP as quantum computing problems have been successful, thus more research on how MT and NLP in general can be re-framed as a quantum computing problem is necessary.

The increasing quality of MT and the expanding preference (especially among younger users) for voice-based interaction with devices points to more and more applications for **speech-to-speech translation and multi-modal machine translation**. Speech translation is a key technique to break the language barrier for human communication. In order to achieve human-like language processing capabilities, machines should be able to jointly process multimodal data, and not just text, images, or speech in isolation. There is a growing need for the translation of audiovisual content and development of MT-centric text-to-speech and speech-to-text applications that can support the meaningful integration of the written and spoken word and images. There is also a need for accessible content in the form of subtitles and audio descriptions. One step in this direction is represented by the recommendations of the New European Media Strategic Research Agenda to develop tools for automatic translation from speech to subtitles, from text to Sign Language, and from Sign Language to text (New European Media Initiative, 2020).

Collection of usable language data is particularly important: while the intensive use of MT systems developed by large global companies allows them to collect and re-use user data, services in Europe would not be able to re-use user data in this way due to GDPR (Aldabe et al., 2021). **Copyright laws** pose a further barrier in Europe: while copyright law is subject to fair-use exceptions in countries such as the US, European law is far less flexible. If lawmakers could agree that using aligned translations of copyrighted data constitutes fair use, as far as it in no way impairs the value of the materials and does not curtail the profits reasonably expected by the owner, LT stakeholders could avail of this high-quality language data.

There is also a disparity between publicly available and proprietary bilingual corpora. Although the EU Coordinated Plan on Artificial Intelligence has foreseen a framework for the free flow of non-personal data in the European Union, the data resulting from public procurement procedures has the tendency to remain locked up in privately-owned data silos, while the research community and LT industry must again to find, identify and reconstruct the public part of this data. A crucial breakthrough could be achieved if existing policy frameworks were adapted to make it mandatory for Member States to make all data in natural language-related workflows publicly available.

Finally, in general, availability and quality of training and test data should be increased. Publicly available multilingual data should include a greater **diversity of domains and languages**, so that building high-quality MT systems becomes an option for all. Future systems should be able to cover all European languages as well as language pairs, and be trained on many different domains and genres.

Current evaluation metrics do not essentially reflect actual translation quality in the real world. Future systems should be evaluated by new automatic metrics which represent better approximations of human judgments and also ideally abandon the dependence on human reference translations. Moreover, evaluation should not be carried out on isolated sentences/segments. Adopting a single metric as a standard for measuring MT would possibly allow for a widespread benchmarking of LT across Europe. Increased attention should be paid to the human judgments used for tailoring the automatic metrics, as well as to manual evaluation in general.

There is also **a lack of necessary resources (experts, HPC capabilities, etc.) in Europe** compared to large US and Chinese IT corporations (e.g., Google, OpenAI, Facebook, Baidu, etc.). While in North America and Asia resources can be allocated to only a limited number of languages, in Europe resources must be distributed across a multitude of official and unofficial EU languages. There is also an uneven distribution of resources across countries, regions and languages (Aldabe et al., 2021). Considering the massive infrastructure that is required to train very large state-of-the-art LT systems, Europe starts with a systemic handicap. Europe's strong foundation in research and innovation can compensate for the disadvantage European organisations have with respect to infrastructure, provided that a concerted effort is undertaken in researching the development of new hardware platforms and respective AI training paradigms.

Finally, the hardware on which MT runs must be scaled down. Several approaches to replace GPU-based computing are already under investigation. By ensuring that the capabilities of the hardware are aligned with the needs of MT training and inference models, smaller models would be easy to integrate and use on any device and also be greener by requiring fewer resources.

Since more and more non-language professionals employ MT daily, there is a need for adequate literacy which includes a measured understanding of LT and its capabilities among the general public. The growing interest in MT literacy is already observed in language industry, where linguists less familiar with MT are taught to use MT critically (Bowker, 2021).

From the end-user/localisation service provider perspective, the pricing pressure often arises as a consequence of not taking into account extra factors which make MT post-editing a more complex task than someone unfamiliar with MT might initially think. In addition, some linguists express negative dispositions towards MT and CAT Tools.

One area where the law is perhaps too relaxed is in relation to carbon emissions in the field of AI research and development. Training neural MT engines is resource intensive and has a heavy carbon footprint. The EU has the opportunity to be a pioneer in training and developing green LT by developing efficient models and hardware, as recommended by Strubell et al. (2019a).

At the level of policies/instruments, much more synchronisation of activities between national and international bodies is necessary. An instrument for efficient and homogeneous implementation of policies towards DLE, would be more equal support for all EU languages, including equal involvement of national research communities.

References

Karim Ahmed, Nitish Shirish Keskar, and Richard Socher. Weighted transformer network for machine translation. *arXiv preprint arXiv:1711.02132*, 2017.

Nur Ahmed and Muntasir Wahed. The de-democratization of ai: Deep learning and the compute divide in artificial intelligence research. arXiv preprint arXiv:2010.15581, 2020.

Itziar Aldabe, Georg Rehm, and Andy Way. Report on existing strategic documents and projects in lt/ai, 2021. URL https://european-language-equality.eu/wp-content/uploads/2021/05/ELE___Deliverable_D3_1.pdf.

Antonios Anastasopoulos and David Chiang. Tied multitask learning for neural speech translation. *arXiv preprint arXiv:1802.06655*, 2018.

Anders Andrae. Total consumer power consumption forecast. *Nordic Digital Business Summit*, 10:69, 2017.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

Doug Arnold, Dave Moffat, Louisa Sadler, and Andrew Way. Automatic test suite generation. *Machine Translation*, 8:29–38, 1993.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, 2017.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. Unsupervised statistical machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3632–3642, 2018.

Srinivasan Arunachalam, Vlad Gheorghiu, Tomas Jochym-O'Connor, Michele Mosca, and Priyaa Varshinee Srinivasan. On the robustness of bucket brigade quantum ram. *New Journal of Physics*, 17(12): 123010, 2015.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural Machine Translation by Jointly Learning to Align and Translate. *CoRR*, arXiv/1409.0473, 2014a. URL http://arxiv.org/abs/1409.0473.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014b.

Sameer Bansal, Herman Kamper, Karen Livescu, Adam Lopez, and Sharon Goldwater. Pre-training on high-resource speech recognition improves low-resource speech-to-text translation. *arXiv preprint arXiv:1809.01431*, 2018.

Yehoshua Bar-Hillel. The present status of automatic translation of languages. *Advances in computers*, 1:91–163, 1960.

Loïc Barrault, Ondřej Bojar, Marta R Costa-Jussa, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, et al. Findings of the 2019 conference on machine translation (wmt19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, 2019.

Lawrence W Barsalou, W Kyle Simmons, Aron K Barbey, and Christine D Wilson. Grounding conceptual knowledge in modality-specific systems. *Trends in cognitive sciences*, 7(2):84–91, 2003.

Hannah Bechara, Raphaël Rubino, Yifan He, Yanjun Ma, and Josef van Genabith. An evaluation of statistical post-editing systems applied to rbmt and smt systems. In *Proceedings of COLING 2012*, pages 215–230, 2012.

Alexandre Bérard, Olivier Pietquin, Christophe Servan, and Laurent Besacier. Listen and translate: A proof of concept for end-to-end speech-to-text translation. *arXiv preprint arXiv:1612.01744*, 2016.

Alexandre Bérard, Laurent Besacier, Ali Can Kocabiyikoglu, and Olivier Pietquin. End-to-end automatic speech translation of audiobooks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6224–6228. IEEE, 2018.

Nicola Bertoldi and Marcello Federico. Domain Adaptation for Statistical Machine Translation with Monolingual Resources. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, pages 182–189, Athens, Greece, March 2009. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W09-0432.

Laurent Bié, Aleix Cerdà-i Cucó, Hans Degroote, Amando Estela, Mercedes García-Martínez, Manuel Herranz, Alejandro Kohan, Maite Melero, Tony O'Dowd, Sinéad O'Gorman, Mārcis Pinnis, Roberts Rozis, Riccardo Superbo, and Artūrs Vasiļevskis. Neural translation for the European Union (NTEU) project. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 477–478, Lisboa, Portugal, November 2020. European Association for Machine Translation. URL https://aclanthology.org/2020.eamt-1.60.

Magdalena Biesialska and Marta R Costa-jussà. Refinement of unsupervised cross-lingual word embeddings. *arXiv preprint arXiv:2002.09213*, 2020.

Steven Bird, Lauren Gawne, Katie Gelbart, and Isaac McAlister. Collecting bilingual audio in remote indigenous communities. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1015–1024, 2014.

Ondřej Bojar and Aleš Tamchyna. Improving Translation Model by Monolingual Data. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 330–336, Edinburgh, Scotland, July 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W11-2138.

Lynne Bowker. Beyond the language industry: Helping others to develop machine translation literacy skills, 2021. URL https://www.gala-global.org/knowledge-center/professional-development/blogs/beyond-language-industry-helping-others-develop.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. Probing the need for visual context in multimodal machine translation. *arXiv preprint arXiv:1903.08678*, 2019.

Erik Cambria and Bebo White. Jumping nlp curves: A review of natural language processing research. *IEEE Computational intelligence magazine*, 9(2):48–57, 2014.

Francisco Casacuberta, Marcello Federico, Hermann Ney, and Enrique Vidal. Recent efforts in spoken language translation. *IEEE Signal Processing Magazine*, 25(3):80–88, 2008.

M. Asunción Castaño, Francisco Casacuberta, and Enrique Vidal. Machine translation using neural networks and finite-state models. In *"Theoretical and Methodological Issues in Machine Translation"*, pages 160–167, 1997. URL http://www.mt-archive.info/TMI-1997-Castano.pdf.

Sheila Castilho. Towards document-level human MT evaluation: On the issues of annotator agreement, effort and misevaluation. In *Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval)*, pages 34–45, Online, April 2021. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2021.humeval-1.4.

Sheila Castilho, Natália Resende, Federico Gaspari, Andy Way, Tony O'Dowd, Marek Mazur, Manuel Herranz, Alex Helle, Gema Ramírez-Sánchez, Víctor Sánchez-Cartagena, Mārcis Pinnis, and Valters Šics. Large-scale machine translation evaluation of the iADAATPA project. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 179–185, Dublin, Ireland, August 2019. European Association for Machine Translation. URL https://aclanthology.org/W19-6732.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4960–4964. IEEE, 2016.

Rajen Chatterjee, Marion Weller, Matteo Negri, and Marco Turchi. Exploring the planet of the apes: a comparative study of state-of-the-art methods for mt automatic post-editing. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 156–161, 2015.

Guanhua Chen, Shuming Ma, Yun Chen, Dongdong Zhang, Jia Pan, Wenping Wang, and Furu Wei. Towards making the most of multilingual pretraining for zero-shot neural machine translation. *arXiv preprint arXiv:2110.08547*, 2021.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D14-1179.

Lonnie Chrisman. Learning recursive distributed representations for holistic computation. *Connection Science*, 3(4):345–366, 1991. URL http://www.chrisman.org/Lonnie/papers/chrisman_cmu_cs_91_154.pdf.

Bob Coecke. The mathematics of text structure. In *Joachim Lambek: The Interplay of Mathematics, Logic, and Linguistics*, pages 181–217. Springer, 2021.

Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. Mathematical foundations for a compositional distributional model of meaning. *arXiv preprint arXiv:1003.4394*, 2010.

Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems*, 32:7059–7069, 2019.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*, 2018.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335, 2017.

Salil Deena, Raymond WM Ng, Pranava Madhyastha, Lucia Specia, and Thomas Hain. Exploring the use of acoustic embeddings in neural machine translation. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 450–457. IEEE, 2017.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Meriam Dhouibi, Ahmed Karim Ben Salem, Afef Saidi, and Slim Ben Saoud. Accelerating deep neural networks implementation: A survey. *IET Computers & Digital Techniques*, 15(2):79–96, 2021.

Mattia Antonino Di Gangi, Matteo Negri, Roldano Cattoni, Roberto Dessi, and Marco Turchi. Enhancing transformer for end-to-end speech-to-text translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 21–31, Dublin, Ireland, 2019. European Association for Machine Translation. URL https://aclanthology.org/W19-6603.

Olivia Di Matteo, Vlad Gheorghiu, and Michele Mosca. Fault-tolerant resource estimation of quantum random-access memories. *IEEE Transactions on Quantum Engineering*, 1:1–13, 2020.

Cem Dilmegani. Current state of ocr: Is ocr dead or is it a solved problem?, 2020. URL https://research.aimultiple.com/ocr-technology/. Last accessed 23 November 2021.

Yanzhuo Ding, Yang Liu, Huanbo Luan, and Maosong Sun. Visualizing and understanding neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1150–1159, 2017.

Zi-Yi Dou, Antonios Anastasopoulos, and Graham Neubig. Dynamic data selection and weighting for iterative back-translation. *arXiv preprint arXiv:2004.03672*, 2020.

Sergey Edunov, Myle Ott, Michael Auli, and David Grangier. Understanding back-translation at scale. *arXiv preprint arXiv:1808.09381*, 2018.

ELRC. Elrc whitepaper: Sustainable language data sharing to support language equality in multilingual europe. Technical report, European Language Resource Coordination, 2019. URL https://lr-coordination.eu/sites/default/files/Documents/ELRCWhitePaper.pdf. more than 70 contributors from research and industry.

GülŞen Eryiğit, Ali Şentaş, and Johanna Monti. Gamified crowdsourcing for idiom corpora construction. *Natural Language Engineering*, pages 1–33, 2021.

European Parliament. Language Equality in the Digital Age. European Parliament resolution of 11 September 2018 on Language Equality in the Digital Age (2018/2028(INI). http://www.europarl.europa.eu/doceo/document/TA-8-2018-0332_EN.pdf, 2018.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833*, 2018.

Manaal Faruqui and Chris Dyer. Improving vector space word representations using multilingual correlation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 462–471, 2014.

Patrick Fernandes, Kayo Yin, Graham Neubig, and André FT Martins. Measuring and increasing context usage in context-aware machine translation. *arXiv preprint arXiv:2105.03482*, 2021.

Mikel L. Forcada and Ramón P. Ñeco. Recursive hetero-associative memories for translation. In *Biological and Artificial Computation: From Neuroscience to Technology*, pages 453–462. Springer, 1997.

Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. Convolutional Sequence to Sequence Learning. *CoRR*, arXiv/1705.03122, 2017. URL http://arxiv.org/abs/1705.03122.

Vittorio Giovannetti, Seth Lloyd, and Lorenzo Maccone. Quantum random access memory. *Physical review letters*, 100(16):160501, 2008.

John J Godfrey, Edward C Holliman, and Jane McDaniel. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pages 517–520. IEEE Computer Society, 1992.

Stephan Gouws and Anders Søgaard. Simple task-specific bilingual word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1386–1390, 2015.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *arXiv preprint arXiv:2106.03193*, 2021.

Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

Jiatao Gu, James Bradbury, Caiming Xiong, Victor O. K. Li, and Richard Socher. Non-Autoregressive Neural Machine Translation. *CoRR*, arxiv/1711.02281, 2017. URL http://arxiv.org/abs/1711.02281.

Çaglar Gülçehre, Orhan Firat, Kelvin Xu, Kyunghyun Cho, Loïc Barrault, Huei-Chi Lin, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. On Using Monolingual Corpora in Neural Machine Translation. *CoRR*, arXiv/1503.03535, 2015. URL http://arxiv.org/abs/1503.03535.

Barry Haddow, Alexandra Birch, and Kenneth Heafield. Machine translation in healthcare. In *The Routledge Handbook of Translation and Health*, pages 108–129. Routledge, 2021.

Jesse Michael Han, Igor Babuschkin, Harrison Edwards, Arvind Neelakantan, Tao Xu, Stanislas Polu, Alex Ray, Pranav Shyam, Aditya Ramesh, Alec Radford, et al. Unsupervised neural machine translation with generative language models only. *arXiv preprint arXiv:2110.05448*, 2021.

Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

Mareike Hartmann, Yova Kementchedjhieva, and Anders Søgaard. Comparing unsupervised word translation methods step by step, 2019.

Bradley Hauer, Garrett Nicolai, and Grzegorz Kondrak. Bootstrapping unsupervised bilingual lexicon induction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 619–624, 2017.

Hiroaki Hayashi, Yusuke Oda, Alexandra Birch, Ioannis Konstas, Andrew Finch, Minh-Thang Luong, Graham Neubig, and Katsuhito Sudoh. Findings of the third workshop on neural generation and translation. *arXiv preprint arXiv:1910.13299*, 2019.

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, 2018.

Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. *Neural Comput.*, 9(8):1735–1780, November 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.8.1735. URL http://dx.doi.org/10.1162/neco.1997.9.8.1735.

Jingjing Huo, Christian Herold, Yingbo Gao, Leonard Dahlmann, Shahram Khadivi, and Hermann Ney. Diving deep into context-aware neural machine translation. In *The Fifth Conference on Machine Translation*, November 2020.

William J. Hutchins. The georgetown-ibm experiment demonstrated in january 1954. In *AMTA*, 2004.

Hirofumi Inaguma, Kevin Duh, Tatsuya Kawahara, and Shinji Watanabe. Multilingual end-to-end speech translation. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 570–577. IEEE, 2019.

Hirofumi Inaguma, Shun Kiyono, Kevin Duh, Shigeki Karita, Nelson Enrique Yalta Soplin, Tomoki Hayashi, and Shinji Watanabe. Espnet-st: All-in-one speech translation toolkit. *arXiv preprint arXiv:2004.10234*, 2020.

Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry P. Vetrov, and Andrew Gordon Wilson. Averaging Weights Leads to Wider Optima and Better Generalization. *CoRR*, arXiv/1803.05407, 2018. URL http://arxiv.org/abs/1803.05407.

Robert A Jacobs, Michael I Jordan, Steven J Nowlan, and Geoffrey E Hinton. Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87, 1991.

Sébastien Jean, Orhan Firat, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. Montreal Neural Machine Translation Systems for WMT'15. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 134–140, Lisbon, Portugal, September 2015. Association for Computational Linguistics. URL http://aclweb.org/anthology/W15-3014.

Wandri Jooste, Rejwanul Haque, and Andy Way. Knowledge distillation: A method for making neural machine translation more efficient. *Information*, 13(2), 2022. ISSN 2078-2489. doi: 10.3390/info13020088. URL https://www.mdpi.com/2078-2489/13/2/88.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. Log-linear combinations of monolingual and bilingual neural machine translation models for automatic post-editing. *arXiv preprint arXiv:1605.04800*, 2016.

Marcin Junczys-Dowmunt and Roman Grundkiewicz. An exploration of neural sequence-to-sequence architectures for automatic post-editing. *arXiv preprint arXiv:1706.04138*, 2017.

Marcin Junczys-Dowmunt, Tomasz Dwojak, and Rico Sennrich. The AMU-UEDIN Submission to the WMT16 News Translation Task: Attention-based NMT Models as Feature Functions in Phrase-based SMT. In *Proceedings of the First Conference on Machine Translation*, pages 319–325, Berlin, Germany, August 2016. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W16-2316.

Daniel Jurafsky and James H. Martin. *Speech and Language Processing (2Nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2009. ISBN 0131873210.

Nal Kalchbrenner and Phil Blunsom. Recurrent Continuous Translation Models. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1700–1709, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D13-1176.

Herman Kamper, Yevgen Matusevych, and Sharon Goldwater. Improved acoustic word embeddings for zero-resource languages using multilingual transfer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1107–1118, 2021.

Takatomo Kano, Sakriani Sakti, and Satoshi Nakamura. Neural machine translation with acoustic embedding. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 578–584. IEEE, 2019.

Dorothy Kenny, editor. *MultiTraiNMT: Machine Translation for Multilingual Citizens*. Language Science Press, Berlin, in preparation.

Yoon Kim and Alexander M. Rush. Sequence-Level Knowledge Distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1317–1327, Austin, Texas, 2016. ACL. URL http://aclweb.org/anthology/D16-1139.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. To Ship or Not to Ship: An Extensive Evaluation of Automatic Metrics for Machine Translation. In *Proceedings of the 6th Conference on Machine Translation (WMT 2021)*, 2021. URL https://arxiv.org/abs/2107.10821. 17pp.

Philipp Koehn et al. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer, 2005.

Taku Kudo. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. *CoRR*, arXiv/1804.10959, 2018. URL http://arxiv.org/abs/1804.10959.

Gaurav Kumar, Matt Post, Daniel Povey, and Sanjeev Khudanpur. Some insights from translating conversational telephone speech. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3231–3235. IEEE, 2014.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*, 2017.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations*, 2018.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. Fully Character-Level Neural Machine Translation without Explicit Segmentation. *CoRR*, arXiv/1610.03017, 2016. URL http://arxiv.org/abs/1610.03017.

J. Lei Ba, J. R. Kiros, and G. E. Hinton. Layer Normalization. *CoRR*, arXiv/1607.06450, July 2016.

Bei Li, Ziyang Wang, Hui Liu, Yufan Jiang, Quan Du, Tong Xiao, Huizhen Wang, and Jingbo Zhu. Shallow-to-deep training for neural machine translation. *arXiv preprint arXiv:2010.03737*, 2020.

Christian Lieske. Metadata and machine translation. *Maschinelle Übersetzung für Übersetzungsprofis. Samelband. BDÜ Fachverlag*, 2020.

Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1871–1880, 2019.

Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. Scheduled sampling based on decoding steps for neural machine translation. *arXiv preprint arXiv:2108.12963*, 2021.

António Lopes, M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins. Document-level neural MT: A systematic comparison. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 225–234, Lisboa, Portugal, November 2020. European Association for Machine Translation. URL https://aclanthology.org/2020.eamt-1.24.

António V Lopes, M Amin Farajian, Gonçalo M Correia, Jonay Trénous, and André FT Martins. Unbabel's submission to the wmt2019 ape shared task: Bert-based encoder-decoder for automatic post-editing. *arXiv preprint arXiv:1905.13068*, 2019.

Ilya Loshchilov and Frank Hutter. SGDR: Stochastic Gradient Descent with Restarts. *CoRR*, arXiv/1608.03983, 2016. URL http://arxiv.org/abs/1608.03983.

Minh-Thang Luong and Christopher Manning. Stanford neural machine translation systems for spoken language domains. In *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, pages 76–79, December 2015. URL http://workshop2015.iwslt.org/downloads/IWSLT_2015_EP_19.pdf.

Minh-Thang Luong and Christopher D. Manning. Achieving Open Vocabulary Neural Machine Translation with Hybrid Word-Character Models. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1054–1063, Berlin, Germany, August 2016. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P16-1100.

Samuel Läubli, Rico Sennrich, and Martin Volk. Has Machine Translation Achieved Human Parity? A Case for Document-level Evaluation. In *Proceedings of EMNLP*, pages 4791–4796, Brussels, Belgium, 2018.

Shuming Ma, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei. Deltalm: Encoder-decoder pre-training for language generation and translation by augmenting pretrained multilingual encoders. *arXiv preprint arXiv:2106.13736*, 2021.

D. Macháček, J. Vidra, and O. Bojar. Morphological and Language-Agnostic Word Segmentation for NMT. *CoRR*, June 2018. URL http://arxiv.org/abs/1806.05482.

Marianna Martindale and Marine Carpuat. Fluency over adequacy: A pilot study in measuring user trust in imperfect MT. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 13–25, Boston, MA, March 2018. Association for Machine Translation in the Americas. URL https://aclanthology.org/W18-1803.

Marianna Martindale, Kevin Duh, and Marine Carpuat. Machine translation believability. In *Proceedings of the First Workshop on Bridging Human–Computer Interaction and Natural Language Processing*, pages 88–95, Online, April 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.hcinlp-1.14.

Sameen Maruf, Fahimeh Saleh, and Gholamreza Haffari. A survey on document-level machine translation: Methods and evaluation. *arXiv preprint arXiv:1912.08494*, 2019.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.448. URL https://aclanthology.org/2020.acl-main.448.

Konstantinos Meichanetzidis, Stefano Gogioso, Giovanni De Felice, Nicolò Chiappori, Alexis Toumi, and Bob Coecke. Quantum natural language processing on near-term quantum computers. *arXiv preprint arXiv:2005.04147*, 2020.

Fandong Meng, Jianhao Yan, Yijin Liu, Yuan Gao, Xianfeng Zeng, Qinsong Zeng, Peng Li, Ming Chen, Jie Zhou, Sifan Liu, et al. Wechat neural machine translation systems for wmt20. *arXiv preprint arXiv:2010.00247*, 2020.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *CoRR*, arXiv/1301.3781, 2013. URL http://arxiv.org/abs/1301.3781.

Azalia Mirhoseini, Anna Goldie, Mustafa Yazgan, Joe Wenjie Jiang, Ebrahim Songhori, Shen Wang, Young-Joon Lee, Eric Johnson, Omkar Pathak, Azade Nazi, et al. A graph placement methodology for fast chip design. *Nature*, 594(7862):207–212, 2021.

Helena Moniz and Carla Parra Escartín, editors. *Towards Responsible Machine Translation: Ethical and legal considerations in Machine Translation*. Springer, Cham, Switzerland, 2022.

Robert C. Moore and William Lewis. Intelligent Selection of Language Model Training Data. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 220–224, Uppsala, Sweden, July 2010. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P10-2041.

Joss Moorkens, Antonio Toral, Sheila Castilho, and Andy Way. Translators' perceptions of literary post-editing using statistical and neural machine translation. *Translation Spaces*, 7(2):240—262, 2018.

New European Media Initiative. Stategic research and innovation agenda 2020, 2020. URL https://nem-initiative.org/wp-content/uploads/2020/06/nem-strategic-research-and-innovation-agenda-2020.pdf?x98588.

Hermann Ney. Speech translation: Coupling of recognition and translation. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, volume 1, pages 517–520. IEEE, 1999.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. Facebook fair's wmt19 news translation task submission. *arXiv preprint arXiv:1907.06616*, 2019.

Khanh Nguyen, Hal Daumé III, and Jordan Boyd-Graber. Reinforcement Learning for Bandit Neural Machine Translation with Simulated Human Feedback. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1464–1474, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D17-1153.

Jan Niehues, Roldano Cattoni, Sebastian Stüker, Matteo Negri, Marco Turchi, T Ha, Elizabeth Salesky, Ramon Sanabria, Loic Barrault, Lucia Specia, et al. The iwslt 2019 evaluation campaign. In *16th International Workshop on Spoken Language Translation*, 2019.

Sharon O'Brien, Patrick Cadwell, and Alicia Zajdel. Communicating covid-19: Translation and trust in ireland's response to the pandemic. Technical report, School of Applied Language and Intercultural Studies, Dublin City University, 2021. URL https://www.dcu.ie/sites/default/files/inline-files/covid_report_compressed.pdf.

Shinhyeok Oh, Sion Jang, Hu Xu, Shounan An, and Insoo Oh. Netmarble ai center's wmt21 automatic post-editing shared task submission. *arXiv preprint arXiv:2109.06515*, 2021.

Lee J O'Riordan, Myles Doyle, Fabio Baruffa, and Venkatesh Kannan. A hybrid classical-quantum workflow for natural language processing. *Machine Learning: Science and Technology*, 2(1):015011, 2020.

Shruti Palaskar, Ramon Sanabria, and Florian Metze. End-to-end multimodal speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5774–5778. IEEE, 2018.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of ACL 2002*, pages 311–318, Philadelphia, Pennsylvania, 2002.

Mārcis Pinnis, Toms Bergmanis, Kristīne Metuzāle, Valters Šics, Artūrs Vasiļevskis, and Andrejs Vasiļjevs. A Tale of Eight Countries or the EU Council Presidency Translator in Retrospect. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track)*, pages 525–546, Virtual, October 2020. Association for Machine Translation in the Americas. URL https://aclanthology.org/2020.amta-user.25.

Marcis Pinnis, Stephan Busemann, Arturs Vasilevskis, and Josef van Genabith. The german eu council presidency translator. *KI - Künstliche Intelligenz*, 2021.

Martin Popel. *Machine Translation Using Syntactic Analysis*. PhD thesis, MFF UK, Praha, Czechia, 2018.

Martin Popel and Ondřej Bojar. Training Tips for the Transformer Model. *The Prague Bulletin of Mathematical Linguistics*, 110:43–70, April 2018. ISSN 0032-6585. doi: 10.2478/pralin-2018-0002. URL https://ufal.mff.cuni.cz/pbml/110/art-popel-bojar.pdf.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtský. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature Communications*, 11(4381):1–15, 2020. ISSN 2041-1723.

Maja Popović. Relations between comprehensibility and adequacy errors in machine translation output. In *Proceedings of the 24th Conference on Computational Natural Language Learning*, pages 256–264, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.conll-1.19. URL https://aclanthology.org/2020.conll-1.19.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018. URL https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf. 12pp.

Aarne Ranta. Explainable machine translation with interlingual trees as certificates. In *Proceedings of the Conference on Logic and Machine Learning in Natural Language (LaML 2017)*, pages 63–78. Gothenburg, 2017.

Spencer Rarrick, Chris Quirk, and Will Lewis. MT Detection in Web-Scraped Parallel Corpora. In *Proceedings of MT Summit XIII*. Asia-Pacific Association for Machine Translation, September 2011. URL https://www.microsoft.com/en-us/research/publication/mt-detection-in-web-scraped-parallel-corpora/.

Georg Rehm and Hans Uszkoreit, editors. *META-NET White Paper Series: Europe's Languages in the Digital Age*, 32 volumes on 31 European languages, Heidelberg etc., 2012. Springer.

Georg Rehm and Hans Uszkoreit. Meta-net white paper series: Press release, 2013. URL http://www.meta-net.eu/whitepapers/press-release.

Natália Resende and Andy Way. Can google translate rewire your l2 english processing? *Digital*, 1(1):66–85, 2021. ISSN 2673-6470. doi: 10.3390/digital1010006. URL https://www.mdpi.com/2673-6470/1/1/6.

Aurko Roy, Ashish Vaswani, Arvind Neelakantan, and Niki Parmar. Theory and Experiments on Vector Quantized Autoencoders. *CoRR*, arXiv/1805.11063, 2018. URL http://arxiv.org/abs/1805.11063.

Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. How2: a large-scale dataset for multimodal language understanding. *arXiv preprint arXiv:1811.00347*, 2018.

Mike Schuster and Kaisuke Nakajima. Japanese and korean voice search. In *ICASSP*, pages 5149–5152, 2012. URL https://research.google.com/pubs/archive/37842.pdf.

Holger Schwenk. Investigations on Large-Scale Lightly-Supervised Training for Statistical Machine Translation. In *Proceedings of IWSLT*, pages 182–189, 2008. URL https://www.isca-speech.org/archive/iwslt_08/papers/slt8_182.pdf.

Rico Sennrich. Why the time is ripe for discourse in machine translation. In *Second Workshop on Neural Machine Translation and Generation*, 2018.

Rico Sennrich, Barry Haddow, and Alexandra 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)*, pages 86–96, Berlin, Germany, August 2016a. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P16-1009.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of ACL 2016*, pages 1715–1725, Berlin, Germany, August 2016b. ACL. URL http://www.aclweb.org/anthology/P16-1162.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Edinburgh Neural Machine Translation Systems for WMT16. In *Proceedings of the First Conference on Machine Translation*, pages 371–376, Berlin, Germany, August 2016c. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W16-2323.

Rico Sennrich, Alexandra Birch, Anna Currey, Ulrich Germann, Barry Haddow, Kenneth Heafield, Antonio Valerio Miceli Barone, and Philip Williams. The University of Edinburgh's Neural MT Systems for WMT17. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 389–399, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W17-4739.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-Attention with Relative Position Representations. *CoRR*, arXiv/1803.02155, 2018. URL http://arxiv.org/abs/1803.02155.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Noam Shazeer, Zhenzhong Lan, Youlong Cheng, Nan Ding, and Le Hou. Talking-heads attention. *arXiv preprint arXiv:2003.02436*, 2020.

Michel Simard, Cyril Goutte, and Pierre Isabelle. Statistical phrase-based post-editing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 508–515, 2007.

Samuel L Smith, David HP Turban, Steven Hamblin, and Nils Y Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *arXiv preprint arXiv:1702.03859*, 2017.

Lucia Specia, Stella Frank, Khalil Sima'An, and Desmond Elliott. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553, 2016.

Lucia Specia, Loic Barrault, Ozan Caglayan, Amanda Duarte, Desmond Elliott, Spandana Gella, Nils Holzenberger, Chiraag Lala, Sun Jae Lee, Jindrich Libovicky, et al. Grounded sequence to sequence transduction. *IEEE journal of selected topics in signal processing*, 14(3):577–591, 2020.

Matthias Sperber and Matthias Paulik. Speech translation and the end-to-end promise: Taking stock of where we are. *arXiv preprint arXiv:2004.06358*, 2020.

Matthias Sperber, Graham Neubig, Jan Niehues, and Alex Waibel. Attention-passing models for robust and data-efficient end-to-end speech translation. *Transactions of the Association for Computational Linguistics*, 7:313–325, 2019.

Felix Stahlberg, Danielle Saunders, and Bill Byrne. An operation sequence model for explainable neural machine translation. *arXiv preprint arXiv:1808.09688*, 2018.

STOA. Language equality in the digital age – Towards a Human Language Project. STOA study (PE 598.621), IP/G/STOA/FWC/2013-001/Lot4/C2, March 2017. Carried out by Iclaves SL (Spain) at the request of the Science and Technology Options Assessment (STOA) Panel, managed by the Scientific Foresight Unit (STOA), within the Directorate-General for Parliamentary Research Services (DG EPRS) of the European Parliament, March 2017. http://www.europarl.europa.eu/stoa/.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy, 2019a.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. *arXiv preprint arXiv:1906.02243*, 2019b.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014. URL https://papers.nips.cc/paper/5346-sequence-to-sequence-learning-with-neural-networks.pdf.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*, 2020.

Pasi Tapanainen and Atro Voutilainen. Tagging accurately–don't guess if you know. *arXiv preprint cmp-lg/9408009*, 1994.

Antonio Toral. Post-editese: an exacerbated translationese. In *Proceedings of Machine Translation Summit XVII: Research Track*, page 273—281, Dublin, Ireland, August 2019. European Association for Machine Translation.

Antonio Toral and Andy Way. What level of quality can neural machine translation attain on literary text? In Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty, editors, *Translation Quality Assessment: From Principles to Practice*, page 263—287. Springer, Cham, Switzerland, 2018.

Antonio Toral, Sheila Castilho, Ke Hu, and Andy Way. Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation. In *Proceedings of WMT*, pages 113–123, Brussels, Belgium, 2018.

Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. Facebook AI WMT21 news translation task submission. *CoRR*, abs/2108.03265, 2021a. URL https://arxiv.org/abs/2108.03265.

Chau Tran, Shruti Bhosale, James Cross, Philipp Koehn, Sergey Edunov, and Angela Fan. Facebook ai wmt21 news translation task submission. *arXiv preprint arXiv:2108.03265*, 2021b.

Joachim Utans. Weight Averaging for Neural Networks and Local Resampling Schemes. In *Proceedings of AAAI-96 Workshop on Integrating Multiple Learned Models*, pages 133–138, 06 1996. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.55.7218&rep=rep1&type=pdf.

Eva Vanmassenhove and Andy Way. SuperNMT: Neural machine translation with semantic supersenses and syntactic supertags. In *Proceedings of ACL 2018, Student Research Workshop*, pages 67–73, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-3010. URL https://aclanthology.org/P18-3010.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 222–232, Dublin, Ireland, August 2019. European Association for Machine Translation. URL https://aclanthology.org/W19-6622.

Eva Vanmassenhove, Dimitar Shterionov, and Matthew Gwilliam. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, page 2203—2213, online, 2021.

Andrejs Vasiljevs, Khalid Choukri, Luc Meertens, and Stefania Aguzzi. Final study report on CEF Automated Translation value proposition in the context of the European LT market/ecosystem, 2019. DOI 10.2759/142151. A study prepared for the European Commission, DG Communications Networks, Content & Technology by Crosslang, Tilde, ELDA, IDC.

Andrejs Vasiļjevs, Inguna Skadiņa, Indra Sāmīte, Kaspars Kauliņš, Ēriks Ajausks, Jūlija Meļņika, and Aivars Bērziņš. Competitiveness analysis of the European machine translation market. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 1–7, Dublin, Ireland, August 2019. European Association for Machine Translation. URL https://aclanthology.org/W19-6701.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6000–6010. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

Ivan Vulic and Anna-Leena Korhonen. On the role of seed lexicons in learning bilingual word embeddings, 2016.

A. Waibel, A. N. Jain, A. E. McNair, H. Saito, A.G. Hauptmann, and J. Tebelskis. JANUS: A Speech-to-Speech Translation System using Connectionist and Symbolic Processing Strategies. In *Proceedings of the 1991 International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 793–796, 1991.

Shuo Wang, Zhaopeng Tu, Shuming Shi, and Yang Liu. On the inference calibration of neural machine translation. *arXiv preprint arXiv:2005.00963*, 2020.

Andy Way. Traditional and emerging use-cases for machine translation. In *Proceedings of Translating and the Computer*, volume 35, London, 2013. 12pp.

Andy Way. Quality expectations of machine translation. In Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty, editors, *Translation Quality Assessment: From Principles to Practice*, pages 159–178. Springer, Cham, Switzerland, 2018.

Andy Way, Petra Bago, Jane Dunne, Federico Gaspari, Andre Kåsen, Gauti Kristmannsson, Helen McHugh, Jon Arild Olsen, Dana Davis Sheridan, Páraic Sheridan, and John Tinsley. Progress of the PRINCIPLE project: Promoting MT for Croatian, Icelandic, Irish and Norwegian. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 465–466, Lisboa, Portugal, November 2020. European Association for Machine Translation. URL https://aclanthology.org/2020.eamt-1.54.

Ron J Weiss, Jan Chorowski, Navdeep Jaitly, Yonghui Wu, and Zhifeng Chen. Sequence-to-sequence models can directly translate foreign speech. *arXiv preprint arXiv:1703.08581*, 2017.

Nathan Wiebe, Ashish Kapoor, and Krysta Svore. Quantum algorithms for nearest-neighbor methods for supervised and unsupervised learning. *arXiv preprint arXiv:1401.2142*, 2014.

Benfeng Xu, Licheng Zhang, Zhendong Mao, Quan Wang, Hongtao Xie, and Yongdong Zhang. Curriculum learning for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6095–6104, 2020.

Hao Yang, Minghan Wang, Daimeng Wei, Hengchao Shang, Jiaxin Guo, Zongyao Li, Lizhi Lei, Ying Qin, Shimin Tao, Shiliang Sun, et al. Hw-tsc's participation at wmt 2020 automatic post editing shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 797–802, 2020.

Jian Yang, Shuming Ma, Haoyang Huang, Dongdong Zhang, Li Dong, Shaohan Huang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, et al. Multilingual machine translation systems from microsoft for wmt21 shared task. *arXiv preprint arXiv:2111.02086*, 2021.

Shaowei Yao and Xiaojun Wan. Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.400. URL https://aclanthology.org/2020.acl-main.400.

Kyra Yee, Nathan Ng, Yann N Dauphin, and Michael Auli. Simple and effective noisy channel modeling for neural machine translation. *arXiv preprint arXiv:1908.05731*, 2019.

Wlodek Zadrozny. On compositional semantics. In *COLING 1992 Volume 1: The 15th International Conference on Computational Linguistics*, 1992.

William Zeng and Bob Coecke. Quantum algorithms for compositional natural language processing. *arXiv preprint arXiv:1608.01406*, 2016.

Xianfeng Zeng, Yijin Liu, Ernan Li, Qiu Ran, Fandong Meng, Peng Li, Jinan Xu, and Jie Zhou. Wechat neural machine translation systems for wmt21. *arXiv preprint arXiv:2108.02401*, 2021.

Biao Zhang, Deyi Xiong, and Jinsong Su. Accelerating Neural Transformer via an Average Attention Network. *CoRR*, arXiv/1805.00631, 2018a. URL http://arxiv.org/abs/1805.00631.

Biao Zhang, Deyi Xiong, and Jinsong Su. Accelerating neural transformer via an average attention network. *arXiv preprint arXiv:1805.00631*, 2018b.

Biao Zhang, Philip Williams, Ivan Titov, and Rico Sennrich. Improving massively multilingual neural machine translation and zero-shot translation. *arXiv preprint arXiv:2004.11867*, 2020.

Baohang Zhou, Xiangrui Cai, Ying Zhang, and Xiaojie Yuan. An end-to-end progressive multi-task learning framework for medical named entity recognition and normalization. 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)*, pages 6214–6224, 2021.

Appendix

A. Additional Material on Transformer

A.1. Attention details

Transformer uses a *scaled dot-product attention*:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{1}$$

where $Q \in \mathbb{R}^{n\times d_k}, K \in \mathbb{R}^{n\times d_k}, V \in \mathbb{R}^{n\times d_v}$, n is the sentence length, d_v is the dimension of values, and d_k is the dimension of the queries and keys.

A.2. Multi-head attention

It is possible to use a single self-attention function in each layer, but the translation quality is improved (Vaswani et al., 2017) when combining multiple *attention heads*:

$$MultiHead(\hat{Q}, \hat{K}, \hat{V}) = \left[head_1(\hat{Q}, \hat{K}, \hat{V}), \ldots, head_h(\hat{Q}, \hat{K}, \hat{V})\right] W^O,$$

$$head_i(\hat{Q}, \hat{K}, \hat{V}) = Attention(\hat{Q}W_i^Q, \hat{K}W_i^K, \hat{V}W_i^V),$$

where h is the number of heads; W_i^Q, W_i^K, W_i^V are parameter matrices which project original size (d_{model}) queries, keys and values ($\hat{Q}, \hat{K}, \hat{V}$) into smaller-size vectors Q, K, V; and $W^O \in \mathbb{R}^{hd_v\times d_{model}}$ is a matrix which projects the concatenation of attention heads back to the original dimension d_{model}. Usually, the "smaller" dimensions d_v and d_k are set both to $\frac{d_{model}}{h}$, but other configurations are possible as well.

A.3. Encoder details

The encoder of Transformer consists of 6 stacked layers of identical form:

$$layer(x) = LN\Big(x + PFFN\big(LN(x + MultiHead(x, x, x))\big)\Big),$$

where $x \in \mathbb{R}^{n \times d_{model}}$ is the input matrix; n is the input sequence length; LN is the *layer normalization* (Lei Ba et al., 2016); $MultiHead$ is the multi-head self-attention sublayer described above; and $PFFN$ is a position-wise feed-forward network (applied on each position independently, thus easy to parallelize):

$$PFFN([x_1, \ldots, x_n]) = [FFN(x_1), \ldots, FFN(x_n)]$$
$$FFN(x_i) = max(0,\, x_i W_1 + b_1)W_2 + b_2$$

In the Transformer "BASE" model, $d_{model} = 512$, $h = 8$ and $d_k = d_v = 512/8 = 64$. In the "BIG" model, $d_{model} = 1024$, $h = 16$ and $d_k = d_v = 64$.

A.4. Positional encoding details

Transformer encodes the absolute position ($pos \in \{1 \ldots n\}$) in the sequence into *positional encoding* vector $PE \in \mathbb{R}^{n \times d_{model}}$.

$$PE_{(pos,2i)} = sin\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$
$$PE_{(pos,2i+1)} = cos\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$