# Chapter 40
# Deep Dive Machine Translation

Inguna Skadiņa, Andrejs Vasiļjevs, Mārcis Pinnis, Aivars Bērziņš, 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, and Andy Way

**Abstract** Machine Translation (MT) is one of the oldest language technologies having been researched for more than 70 years. However, it is only during the last decade that it has been widely accepted by the general public, to the point where in many cases it has become an indispensable tool for the global community, supporting communication between nations and lowering language barriers. Still, there remain major gaps in the technology that need addressing before it can be successfully applied in under-resourced settings, can understand context and use world knowledge. This chapter 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 if the goal of digital language equality in Europe is to be achieved.[1]

Inguna Skadiņa · Andrejs Vasiļjevs · Aivars Bērziņš · Mārcis Pinnis
Tilde, Latvia, inguna.skadina@tilde.com, andrejs.vasiljevs@tilde.com, aivars.berzins@tilde.com, marcis.pinnis@tilde.com

Nora Aranberri · Iakes Goenaga
University of the Basque Country, Spain, nora.aranberri@ehu.eus, iakes.goenaga@ehu.eus

Joachim Van den Bogaert
CrossLang, Belgium, joachim.van.den.bogaert@crosslang.com

Sally O'Connor · Riccardo Superbo
KantanMT, Ireland, sallyoc@kantanai.io, riccardos@kantanai.io

Mercedes García-Martínez · Manuel Herranz
PANGEANIC, Spain, m.garcia@pangeanic.com, m.herranz@pangeanic.com

Jan Hajič · Martin Popel · Rudolf Rosa
Charles University, Czech Republic, hajic@ufal.mff.cuni.cz, popel@ufal.mff.cuni.cz, rosa@ufal.mff.cuni.cz

Christian Lieske
SAP SE, Germany, christian.lieske@sap.com

Maja Popović · Sheila Castilho · Federico Gaspari · Andy Way
Dublin City University, ADAPT Centre, Ireland, maja.popovic@adaptcentre.ie, sheila.castilho@adaptcentre.ie, federico.gaspari@adaptcentre.ie, andy.way@adaptcentre.ie

[1] This chapter is an abridged version of Bērziņš et al. (2022).



G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_40

# 1 Introduction

Machine translation (MT) was one of the first application areas of natural language processing (NLP). Starting from the first attempts to apply dictionary-based approaches right up to modern neural network-based 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 languages, breaking down language barriers and supporting the implementation of the European digital single market. For example, the eTranslation automated translation tool,[2] developed by the European Commission, and its various adoptions (e. g., EU Council Presidency Translator, Pinnis et al. 2021)[3] provide reasonably good MT service in 24 EU official languages for governments, the public sector and SMEs.[4] However, MT support and the quality of its output 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 available data.

## 1.1 Scope of this Deep Dive

In 2012, the META-NET White Paper series (Rehm and Uszkoreit 2012) presented a thorough analysis of Language Technology (LT) support for 31 European languages. According to this study, for MT *good support* only applied to English and *moderate support* to only two widely spoken languages (French and Spanish), leaving the remaining 28 European languages in clusters of *fragmented* or *weak or no support*.

This chapter focuses on the MT landscape a decade after the publication of the META-NET White Papers. We analyse progress in MT, identify the main gaps and outline visions, the breakthroughs needed and development goals towards Digital Language Equality (DLE) and Deep Natural Language Understanding (NLU) by 2030. We look at the current services and technologies offered by MT providers in the European market. The dominance of global companies in the free online translation market and the risks for Europeans caused by this dependence are among the key topics discussed in this chapter, especially to identify solutions going forward.

The main gaps are identified for four dimensions of MT: data, technology, approaches and legislation. We focus not only on data availability and usability and the need for less-resourced technologies, but also discuss limitations related to multimodal MT. While MT technologies today are available for most European languages, many of these languages are less attractive from a business point of view, and con-

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

[3] https://www.eu2020.de/eu2020-en/presidency/uebersetzungstool/2361002

[4] As of February 2022, eTranslation was used by 108 projects – 87 projects reusing eTranslation and 21 projects committed to analysing or reusing eTranslation.

sequently they are not so well equipped with MT tools. Throughout the chapter, language coverage is addressed as a key dimension for DLE. We also discuss legal and ethical aspects related to the development, production and use of MT systems and services. We analyse 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 these aspects are taken into consideration from the perspective of their impact on society, with a focus on Europe. The chapter provides a series of recommendations on how to address the current limitations of MT technologies and how to contribute to DLE as a crucial goal for Europe and its citizens.

## 1.2 Main Components

While different MT types (e. g., rule-based, example-based, statistical, hierarchical) have been investigated, in this subsection we will focus only on the recent development of Neural MT (NMT), based on an overview by Popel (2018). We present the main MT components of the general NMT architecture and the currently most popular example: Transformer (Vaswani et al. 2017). There are many other components related to MT, which are not described here, e. g., automatic speech recognition[5] and speech synthesis, which are needed in the speech-to-speech translation pipeline; cross-lingual information retrieval; multilingual summarisation; integration into production systems and multilingual websites using suitable metadata formats.[6]

In NMT, each input sentence is first tokenised into a sequence of tokens. The most popular approach today is to split words into subword units (subwords, which need not be actual words of the language or even morphemes). For example, the German word *Forschungsinstituten* ('research institutes') may be encoded with three subwords: *Forsch* + *ungsinstitu* + *ten_*. There are several algorithms for training subword models (e. g., Sennrich et al. 2016b). NMT based on subwords shows better results than early approaches based on words and recent approaches based on characters (Libovický et al. 2022). Each token is represented as a real-value vector, called (subword/word) embedding. Most NMT systems initialise embeddings randomly and train them jointly with the whole translation, but pre-trained (contextual) embeddings may be used as well, especially in low-resource settings.

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. 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.

[5] See, for example, the reports of the ELITR project at https://elitr.eu.

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

(2015) 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 translation quality, especially on longer sentences.

The process of translating sentences (at test time) with a trained NMT model is usually called inference. Most NMT systems use auto-regressive inference. This means that the output sentence is generated token by token and after each token is generated, its embedding is used as input for generating the next token. Decoding finishes once the decoder generates a special end-of-sentence token.

The advantage of NMT systems is that all their components can be trained in an end-to-end fashion unlike earlier data-driven approaches, 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; there are also NMT systems optimising sentence-level metrics (e. g., BLEU, Papineni et al. 2002, or simulated human feedback) with reinforcement learning techniques (e. g., Nguyen et al. 2017). NMT usually uses teacher-forcing: when generating the next word during training, it uses the previous word from the reference translation as the input instead of using the previously predicted word.

The Transformer architecture follows the general encoder-decoder architecture, but unlike earlier recurrent-networks it uses self-attention and feed-forward layers in both the encoder and decoder. This allows training and partially also the decoding process to be sped up thanks to better use of parallelisation.

Self-attention is based on a compatibility function which assigns a weight to each pair of tokens, more precisely, to their vector representation on each layer. Transformer uses multi-head self-attention, so multiple versions (heads) of the self-attention function are trained for each layer. Figure 1 shows an example of visualisation for different heads.

## 2 State-of-the-Art and Main Gaps

### 2.1 State-of-the-Art

Deep learning techniques have given a major boost to the area. The application of neural networks to MT has opened the path to developing a universal engine whose ultimate goal is a single model to translate between any arbitrary language pair. The effects of different advanced approaches for multilingual MT models have been investigated by Yang et al. (2021), for example. They first explore how to leverage the large-scale language models created from 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 (e. g., Zhang et al. 2020) to create a deep model from a shallow one. Additionally, they implement multiple rounds of back-translation (e. g., Dou et al. 2020) for data augmentation purposes. While the

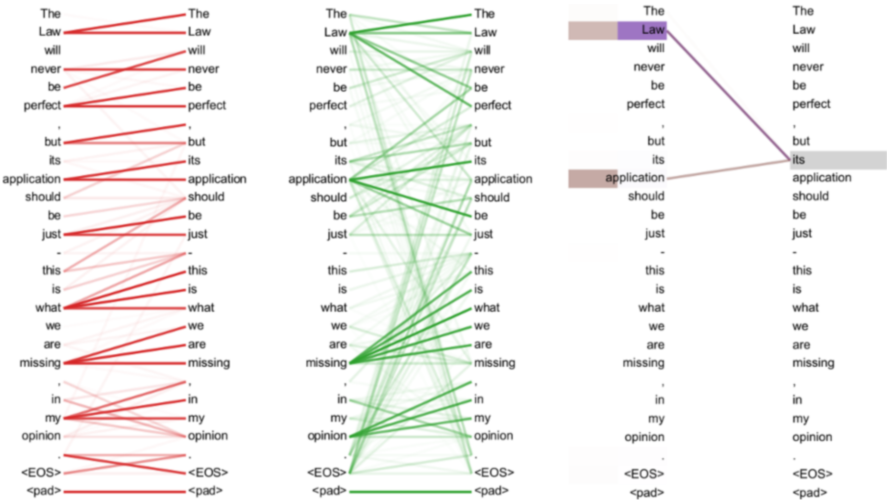


**Fig. 1** Visualisation of self-attention in a Transformer model trained on English → German translation (adapted from Vaswani et al. 2017). Each head is visualised in a different colour and edge weight is indicated by thickness. Each of the figures shows another attention head in encoder layer 5 (out 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. The right-most figure shows two attention heads, but focusing only on the word '*its*' and illustrating coreference resolution.

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. However, the results worsen considerably when translation in language pairs without English is considered.

If we turn to the goal of achieving language equality, one of the most interesting approaches is unsupervised MT (e. g., Artetxe et al. 2018) where no bilingual parallel data is needed to train a fully working system. In recent years, this approach has slowly been catching up with the translation quality obtained by supervised systems. For instance, Han et al. (2021) build a state-of-the-art unsupervised NMT system derived from a generative 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) 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, which is then distilled into a smaller model via fine-tuning to obtain a new state-of-the-art in unsupervised translation on the WMT14 English-French benchmark. While still restricted to a well-resourced language pair, learning outcomes are promising for lower-resource pairs.

Within the industrial context, a look at providers' solutions gives a clear overview of the strengths of each company, as well as the issues that remain relevant regarding the successful implementation of the technology. A key aspect that most companies emphasise is the capacity for domain adaptation. This allows for engines that learn from domain-specific texts, avoiding the noise that expressions from other fields might introduce in the learning process (e. g., Pangeanic, RWS, Tilde, Welocalize). Further customisation is also highly valued, most frequently by refining their own generic or domain-specific engine with a customer's own data (e. g., Across, Language I/O, Tilde). Alternatively, do-it-yourself MT opportunities are provided where customers build their own system from scratch using just their own data.

The text type involved is also distinctive across companies, with some pushing for real-time adaptive MT for email and chat (e. g., Language I/O), while others emphasise multimodality. When a level of accuracy and/or cultural adaptation is required, MT is coupled with post-editing, which is implemented with functionalities directed at professional translators or crowd-sourcing platforms (e. g., Lengoo, Unbabel).

Apart from the quality of the technology itself, seamless integration within existing localisation workflows is paramount for its successful adoption, as well as scalability (e. g., KantanMT, Lilt, Tilde), open-source technology (e. g., Pangeanic, Apertium) and speech MT (e. g., Papercup, Tilde). Additionally, privacy and security are of huge interest as texts often include sensitive product or customer information. The lack of understanding of how MT works and the unclear legal rights, obligations and consequences of misuse cause clients to seek secure solutions (e. g., Across, Language Weaver, Pangeanic, Tilde).

There are numerous European companies providing MT tools and services, each with their own strengths and limitations. However, it is tech giants such as Amazon, Facebook, Google, Microsoft who set the standards and best practices for LT development and provision. Most such companies are headquartered outside Europe and so 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 the European level results in a disjointed European market with disparate support for the language communities of Europe. Such a roadmap is crucially important now that MT is playing a key role in communication activities across the globe. As a result, the demand for translated content has reached an all-time high, but seems set to rise for the foreseeable future.

Nowadays, there are countless online MT sites for general use that offer access to MT either from companies that make the systems freely available with some usage restrictions (Amazon, Google, Microsoft, DeepL, and Tilde among others) or from public bodies that facilitate their custom-based MT capabilities (the European Commission and the Basque and Latvian governments, among others, Skadins et al. 2020). People use these tools to translate a very diverse range of texts. While access is fast and straightforward, they do present privacy risks and cultural bias. To this day, the legal boundaries of text ownership and use are not fully regulated across Europe. Also, the array of languages available is increasing, but it is the major languages that benefit from the advances first and foremost, with small and minority languages often suffering from uneven and generally low quality.

MT has been available to the video game localisation industry for years without much success given the need for highly creative and culturally adapted options, often with constraints dictated, for example, by available on-screen space. For current online collaborative games, in-game dialogue has become critical, as has the need for instant translation between multiple languages. This has motivated some game developers to explore the potential of MT in their localisation processes.

Medical translation is highly sensitive and requires the utmost precision. Given the serious consequences of mistranslations, MT has been largely absent from this area. However, it is time to push for MT accuracy and consistency, and accept nothing short of high-quality translation (Haddow et al. 2021). MT could prove of great assistance not only for written text but also in doctor-patient communication. While medical interpreters remain the go-to specialists, often their services are not available. To facilitate this type of communication, systems that can specifically tackle the local languages and those of the immigrants are essential. There are now a number of success stories that demonstrate the utility of MT in this field. For example, in 2020 SDL made their MT system available to all engaged in COVID-19 medical research;[7] NAVER LABS Europe released an MT model for COVID-19 research;[8] and, to make emergency and crisis-related content available in as many languages as possible, Translators without Borders and several academic and industry partners prepared COVID-19 materials for training MT models for nearly 90 languages.[9]

Public Administration – Making legal and administrative documents available in at least the official languages of Europe is an obligation of national governments. Given the intricacies of the texts, MT is not yet central in the translation process. However, several initiatives such as ELRC[10], ELRI[11] and ELG[12] (Rehm et al. 2023) have curated and shared LRs that can improve MT services. Along the same lines, the availability of high-quality NMT at different levels of public bodies, Member States and public administrations has been put forward as a key priority for the European Commission, particularly for under-resourced EU languages (see, e. g., the projects NTEU and iADAATPA, Bié et al. 2020; Castilho et al. 2019). An excellent example of the use of MT by EU Council Presidency staff members and public administration translators is demonstrated by the eight EU Council presidencies that used the EU Council Presidency Translator (Metuzale et al. 2020). The challenge is the provision of this type of service not only for the 24 official languages, but for all languages in Europe, promoting citizen equality and European cohesion, which are key to a stable and unified view in the region.

To increase customers' understanding of a product and to build trust, global content on an eCommerce website should be translated into the target customer's language. eCommerce companies require a mix of technical, highly accurate yet informal, creative, and culturally aware translations. While that can be challenging for MT, there are many companies (e. g., Lionbridge, Protranslating, Simultrans, Smartling) that can help online business owners to make their content multilin-

7 https://www.biospace.com/article/releases/sdl-offers-machine-translation-free-of-charge-to-health-science-professionals-

8 https://europe.naverlabs.com/blog/a-machine-translation-model-for-covid-19-research

9 https://tico-19.github.io

10 https://www.lr-coordination.eu

11 http://www.elri-project.eu

12 https://www.european-language-grid.eu

gual, with multiple plugins compatible with common Content Management Systems (CMS) and eCommerce solutions in the market (WordPress, Drupal, Joomla, Magento and WooCommerce).

This short review shows that the current shortcomings of MT technology and areas where effort should concentrate revolve around aspects that help increase trust through increased accuracy, as well as 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 scenario where all citizens feel equal, with the same quality of language access to resources, services and commerce, will considerably boost European cohesion.

## 2.2 Main Gaps

*Data Availability and Data Quality* – As stated in the EU Charter and the Treaty on the EU, all 24 official EU languages are granted equal status. However, the META-NET White Paper Series found that 21 of the 30 European languages investigated were at risk of digital extinction. 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: 1. Europeans are not receiving the digital resources they are entitled to; and 2. there is a lack of language data to train MT engines to mitigate this problem. The Open Data Directive (2019/1024/EU) does not recognise language data as a high-value data category. This means that it may not be clear what language data exists for at-risk languages, or how data can be used for MT/LT development. Moreover, availability does not guarantee usability. To be considered usable, language data must meet certain criteria. For instance, to train high-performance NMT systems, bilingual data needs to be clean and correctly aligned.

*Domain-specific Data* – NMT 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 an NMT system. This means that domain availability is almost as relevant as language availability. While categories such as legal, financial, and technical are usually well covered in terms of availability and suitability for a number of languages and language pairs, more specific or uncommon domains may not have comparable amounts of training data available. Moreover, there is generally a disparity between publicly available and proprietary bilingual corpora. As a result, there is a gap in the availability of domain-specific language data both in official and minority languages, which could lead to the centralisation of some specialised fields over others, excluding speakers of less supported languages in the long term.

*The Compute Divide* – With the paradigm shift to NMT, MT has become increasingly computationally intensive. Access to hardware, experts, and involvement in research has also shifted in such a way that elite universities and larger enterprises have an advantage due to their relative ease of access to compute power. According to the ELE analysis on strategic documents and projects (see Chapter 44, p. 361 ff.), there is a lack of necessary resources (experts, High Performance Computing, capabilities, etc.) in Europe compared to large US and Chinese IT corporations that lead the development of new LT systems. Furthermore, there is an uneven distribution of resources, including scientists, experts, computing facilities, and companies, across countries, regions and languages in Europe (cf. Rehm et al. 2023).

*Multimodal MT* – MT is commonly thought of as translating text to text, but multimodal MT is also possible, although it is still in its early stages. Fields in which further technological innovation would increase potential use-cases for MT include image recognition, speech synthesis and automatic speech recognition. Image-to-text translation makes use of Optical Character Recognition (OCR) to isolate text in images. This technology is quite effective, and nowadays smartphone 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 OCR accuracy, including coloured or decorative backgrounds, blurred texts, non-Latin alphabets, larger or smaller letters, look-alike characters, and handwritten text, all or any of which may result in nonsensical translations. Combining OCR with text prediction may improve the accuracy of this technology. Audiovisual media is playing an increasingly central role in our lives thanks to 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 Strategic Research Agenda developed by New European Media[13] provides a number of recommendations related to MT, including 1. streamlining the circulation of audiovisual (or video) programs through MT, while humans focus on the quality of work, for example; 2. encouraging synergies and convergence between subtitling and the development of multilingualism or the integration of foreign migrants, for example; 3. developing AI tools for automatic translation from speech to subtitles, and text to/from sign language; and 4. developing AI tools for robust automatic translation of subtitles. Training high-performance MT systems to translate subtitles is particularly challenging. 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. 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.

*Different Types of End Users* – The language industry is 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 fac-

[13] https://nem-initiative.org

tors, including the quality of the MT output and the complexity of the content or domain. The pricing pressure often arises from a lack of consideration of these extra factors which make post-editing a more complex task than it initially appears. Providing industry with the resources to better communicate these factors could be a step towards relieving pricing pressure. Furthermore, LT has changed the role of the translator.[14] 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. However, linguists play an important role in the assessment and continuous improvement of MT engines, because there is no universal way to automatically evaluate MT quality. Therefore, while the role of traditional translators might have changed, demand for linguists has remained high alongside the developments of MT. At the other end of the spectrum, the hype about the advancements of AI and MT might convince people with low levels of expertise into thinking that MT is infallible (for clear demonstrations that the 'human parity' claims were less than watertight, see Läubli et al. 2018; Toral et al. 2018). The wide availability of MT applications coupled with the sometimes deceptive fluency of NMT output may lead users to avail of MT uncritically, without always understanding its pitfalls. Another step in this direction includes educational publications, which address the technical foundations of machine learning as used in MT as well as the ethical, societal, and professional implications of its use (Kenny 2022).

*Automated Evaluation of MT* – Automated metrics are a cost-effective way of assessing the quality of MT output. Research in the field focuses heavily on developing metrics that are able to show higher and higher correlations with human judgement. As a result, different metrics are presented at conferences around the world every year. Despite (or 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. Adopting a single metric as a standard would possibly allow for a widespread benchmarking of MT across Europe.

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.

*Licensing* – 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.

*Copyright* – 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, and severely restricts the use of parts of copyright works for purposes such as data mining. If lawmakers could agree that using aligned translations of

[14] We use the word *linguist* to refer to language professionals who translate, post-edit, and evaluate LT among other tasks

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 for the immediate benefit of European language communities.

*Legislative and Adoption Gaps* – Despite the widespread celebration of multilingualism in the EU, there is no common policy addressing language barriers as of yet. We now provide a few 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. 1. the COVID-19 pandemic has shown the need for rapid dissemination of information and guidelines in times of crisis. To give one example, in Ireland, the provision of multilingual information was seen to be slow, and reactive, with even the provision of information in Irish and Irish Sign Language being slow in the early stages. The first recommendation made (O'Brien et al. 2021) is for state departments to implement a coordinated approach to the provision of translated content in crises; 2. the requirement for all translations of personal documents to be stamped by a sworn translator can increase the stress on civilians, adding costs and waiting times. The repetitive nature of documents like these as well as their standardised terminology are particularly well-suited to MT; 3. 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 eCommerce; 4. there is a gap in publicly available MT services which cater specifically to the needs of people in Europe. Users can globally avail of free-of-charge MT services but the multinationals who provide the services could withdraw or start charging for them at any time. Moreover, they do not cater specifically to the needs of European citizens.

Training NMT 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. (2019) 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.

# 3 The Future of the Area

In this last section, we will examine the contribution of MT to DLE (Section 3.1), briefly sketch the main breakthroughs needed (Section 3.2), discuss our main technology development goals and visions (Section 3.3) and describe the next steps towards Deep NLU (Section 3.4).

## 3.1 Contribution to Digital Language Equality

Nowadays, due to globalisation, MT is essential for the development of society. People can access MT allowing for the democratisation of information in many languages. MT directly impacts the economy and cultural exchange between countries. In various scenarios, human translators cannot meet the huge demand for translations in a short time and at low cost. In such cases, MT is much faster and may require less effort to post-edit than translating from scratch.

Massive amounts of parallel data are required to build solid MT systems. Parallel data creation is costly in terms of time and resources. We contend that work done for or by public administrations might offer a solution in this regard. The NEC TM project,[15] for example, calculated in its market study that European public administrations spend about 300 million Euros p. a. in translation contracts with language vendors. This parallel data is mostly not requested back by institutions, many of which operate in low-resource languages, but it should be made publicly available. Data availability directly affects the availability and quality of MT, as well as the contribution it can make to DLE and the wider society. These data pipelines can improve local (national) technology, raise awareness of the fact that citizens are also data producers, and improve and increase the availability and quality of MT. For example, in the case of Catalan, having co-official status (in three Spanish regions) kickstarted a series of administrative decisions that facilitated the creation of more and more parallel data, which has been utilised by local MT companies. Societies that care about data sovereignty and establish language data policies can facilitate the growth of LT companies, which in turn can positively impact those societies.

Uses of MT are very varied, from customer reviews on travel sites to legal document translation for public administrations. None of those uses and the business intelligence that can be derived from them can happen without translation. 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 businesses 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 is now embedded in many services as a real-time high-quality commodity. The ELE consortium has identified several day-to-day uses which illustrate how

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

MT is used in very different spheres, including: 1. civil servants verify the national legislation of other EU Member States by machine-translating it; 2. citizens communicate via MT when visiting other countries; 3. the general public use MT to understand social media conversations; 4. students machine-translate research papers; 5. eCommerce websites offer products online to consumers in multiple languages; and 6. public administrations translate documentation for information exchange.

All these use-cases generate massive amounts of online data, that is not reused by EU businesses and research groups. Worse still, it can happen that it is generated for the benefit of the (non-European) free online tools providers to make their technology more accurate. Access to massive amounts of data that is freely available and provided by general users has scaled a lot of MT research, whilst it has provided little in terms of open-source, generally available resources.

Whilst the majority of the talent in NLP and AI has been European, large-scale developments are foreign to the EU or the result of private sponsorship. Heavy investment in MT research at universities over the years has created the know-how and technical knowledge which has only rarely been exploited commercially (e. g., KantanMT, Iconic). The question for Europeans remains on the privacy of the data used and how this data is transmitted. The MT landscape is dominated by large non-European players and technology companies. DeepL is the only significant 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 much less 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 SMEs in 2021.

A good example of increasing concerns comes from Switzerland, where DeepL and Google Translate were recently banned at Swiss Post as external tools amid concerns of privacy and data exploitation (access was later reopened, though). Swiss Post declared 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.[16] GDPR has the potential to change things as privacy concerns become relevant to institutions and enterprises, with EU projects such as MAPA[17] providing 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.

## 3.2 Breakthroughs Needed

According to a competitiveness analysis ordered by the European Commission, 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 behind in terms of in-

16 https://slator.com/swiss-post-bans-deepl-backs-down-after-staff-uproar/

17 https://mapa-project.eu

vestment, infrastructure and industry implementation (Vasiljevs et al. 2019). At the same time, the study highlights that the market is fragmented, which 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 MT/LT systems, Europe starts with a systemic handicap. Looking forward to 2030, we expect the movement towards more efficient and real-time translation to continue. 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 AI training paradigms.

For Europe, a breakthrough in these fields is needed to remain on par with the rest of the world. Breakthroughs in the development of hardware platforms and training paradigms are also warranted by several EU policies. Through the European Green Deal[18] and the Horizon Europe Work Programme (European Commission 2021), the European Commission has committed to making "Europe the world's first climate-neutral continent by 2050", i. e., 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. If we want MT to become ubiquitous, especially in embedded devices, the hardware on which it runs must be scaled down and the models that run on it 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; both need to be developed in cooperation to ensure that the capabilities of the hardware are aligned with the needs of MT training and inference.

An equally fundamental breakthrough is needed in the understanding of how our current algorithms work. Many NLP systems today are based on large pre-trained language models which have demonstrated outstanding results on different tasks. However, a boost in performance comes with a cost in efficiency and interpretability, which "is a major concern in modern Artificial Intelligence and NLP research, as black-box models undermine users' trust in new technologies" (Fomicheva et al. 2021). The EU Coordinated Plan on Artificial Intelligence (ECPAI, European Commission 2018) 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, in order to prevent models from performing poorly or from acting in a gender- or culturally-biased manner.

The ECPAI correctly states that "[f]urther developments in AI require a well-functioning data ecosystem built on trust, data availability and infrastructure", but

[18] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52019DC0640

it underestimates the effect that one of its cornerstones has had on data collection in the field. According to the plan, "[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. […] 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." (European Commission 2018).

Unfortunately, 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 GDPR. This is especially true for human language data, since it has no fixed schema indicating when personal details may occur. As a result, expensive legal counsel and 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, which has given them a considerable competitive advantage over EU companies.

Although the ECPAI has foreseen a framework for the free flow of non-personal data in the European Union (European Union 2018b), including the creation of common European data spaces in a number of areas, and a proposal for a directive on the reuse of public sector information (European Union 2018a), the process of obtaining linguistic data that has been created using public funding is currently far too cumbersome and pull-oriented. The data resulting from public procurement procedures has a tendency to remain locked up in privately-owned data silos, while the research community and LT industry must go to great lengths to identify and reconstruct the public part of this data using NLP tools (see, for example, Koehn 2005). 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, 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 service administrators should be made aware of the value of their human workflows. As a second step, the IP resulting from public service workflows should be publicly released by default. Finally, workflow data should be made discoverable in a publication/subscription manner, so it can be easily picked up by interested parties.

Although MT has taken a big leap forward with the advent of neural systems, some types of translation remain very difficult. If we want MT to become pervasive for problematic text types (spreadsheets with tabular data, metadata fields, etc.), the problem of context modelling needs to be addressed. For textual translation, incorporating ontological information may help. Continued development on multilingual lexical resources will be required for this. For multimodal settings, extra-lingual context must be incorporated to improve results. Context modelling is not only required to deal with short sentences or phrases, but also to obtain more cohesive translation across larger volumes of text. NMT systems have improved over SMT, but have not yet succeeded in efficiently incorporating basic grammatical relations between sentences and paragraphs. Since the majority of human language is produced outside of written texts, extra-lingual cues are often required to decode a message adequately and to translate it correctly. To enable better modelling of multimodal 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 modelling.

In terms of the development of data, two important breakthroughs which must be achieved are 1. the creation of new data sets, and reiteration over existing data sets; and 2. policy support for public data reuse. Ideally, new data annotation efforts should build upon existing work. For example, 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 existing NLP technology it is not unthinkable that EU Parliament sessions could 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 multimodal features towards translation quality.

There are various other fields and areas in which further breakthroughs are needed, some of which are novel methods for document-level MT (with a focus on coherent translations of whole texts and documents), the integration of visual and audio features into MT approaches and engines as well as improved explainability (see Bērziņš et al. 2022). Another field is quantum computing, where more research is needed on how MT, and NLP in general, can be reframed as a quantum computing problem. Current work is still laying the foundation for future developments, 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 already been made.

## 3.3 Technology Visions and Development Goals

The strategy of building huge MT models by collecting 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, so smaller models adapted to particular language pairs and domains need to be made available. This would have several benefits: such models would be easy to integrate and use on any device, provide high-quality translations for all domains and languages, and also be greener by requiring fewer computational resources.

The future publicly available MT systems should be less dependent on large companies, especially those which are not European. The risk is that what is freely available now 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.

Another challenge of the current systems is represented by various biases in the models, such as gender, racial and ethnic bias (Vanmassenhove et al. 2019). Such 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.

More and more NMT systems are being developed which go beyond the single sentence level (e. g., Lopes et al. 2020), using a variety of different approaches: taking into account source- or target-language 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'. 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. Future systems should be able to identify which sentences benefit from the availability of context, and then find that context. This task is far from trivial because relevant information can be found in different places, sometimes even beyond the given text, such as the topic of the text, the gender of the writer/speaker, or even general world knowledge.

Such 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 sources of information which go beyond this, so that an image of a product can help disambiguate words in the description or review of the said product, for example. Multimodal models should also include sign language translation, which currently relies mainly on computer vision methods. Sign language MT should use models based on both images and natural language.

Training data, crucial to building models, should receive more attention. 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 progress in MT is mainly measured under 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 – the availability and quality of training data should be increased. Attention should also be given to languages where there is no written tradition, in which case spoken-language data needs to be sourced.

While techniques such as multilingual models, unsupervised MT, synthetic data, and transfer learning are all helping, if there is not enough good-quality data for a language (pair), then such methods will not reach the goal of high-quality MT, in which case novel methods and research breakthrough will be needed in this direction.

The test sets used for assessing MT systems should receive more attention, too. Currently, a large number of research publications use news articles coming from shared tasks. Researchers test their systems on these texts and report improved automatic scores. However, some of the human translations in these test sets used as references for automatic scores are of poor quality (Toral et al. 2018). 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 long-term testing. Note that in light of the comparison between MT outputs and human translations carried out in recent years where claims of "human parity" have been investigated, the quality of human translations used in MT evaluation has to be high (Toral et al. 2018; Läubli et al. 2018).

In addition, other test sets coming from different genres and domains need to be more widely used. A vast amount of systems are currently tested only on a limited set of domains, news being the predominant one, while many genres and domains are as yet hardly covered by current research, such as user-generated content (which itself is not a homogeneous genre), despite having great potential for future growth. In the long run, 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. In this regard, the rise of NMT and its increasing quality have led to more and more challenge test sets (or test suites). These specified test sets enable better understanding of certain (linguistic) aspects which cannot be properly assessed in standard 'natural' test sets. The development and creation of such test sets necessitate a large amount of human expertise, time and effort. In the future, they should be easy and fast to create for any language pair.

As for the evaluation process itself, automatic metrics remain invaluable tools for the rapid development and comparison of MT systems. They have been developed and improved constantly, with more and more metrics coming onstream each year. However, a number of challenges remain. Perhaps the most significant is that the community still relies to a large extent on BLEU, despite there being a large body of research pointing out its drawbacks. Future systems should be evaluated by new metrics which represent better approximations of human judgments and also ideally abandon the dependence on human reference translations, which is a serious limitation. Recently, more and more metrics based on neural networks and/or word representations have emerged which show better correlation with human judgment and do not require reference translations. However, these metrics have another limitation: they require labelled training data which as we have pointed out are available only for a limited number of language pairs and domains. Future automatic metrics should be equally valid without such constraints. In addition, all future automatic metrics should be able to evaluate MT output taking the context into account in order to be more reliable (Läubli et al. 2018; Castilho 2021).

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 and 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. We assert 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, as pointed out by Way (2013) and Mason (2019). These aspects might include adequacy/accuracy, readability/comprehension, appropriate register, correct terminology, or adequately fulfilling a particular task. Consequently, metrics should be created with such criteria 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 (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.

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. 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 needs to be 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 (Gangi et al. 2019), but clearly more work needs to be done in this regard.

Sign language translation should be widely available for many domains to break down language barriers for deaf and hearing-impaired users so that they can access information like the rest of society. For this to be done properly, sign language translation needs to include language features in addition to image features. In addition, it should not only be translated from/into text but also from/into speech.

It is more and more the case that MT is 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 be necessary. Extracting different representations from various layers could be even better suited for certain tasks, so this option should be made easily available in future MT systems.

### 3.4 Towards Deep Natural Language Understanding

Applying a purpose- and communication-oriented view on MT allows us to discuss the extent to which MT needs (deep) NLU, since it helps to put the prevailing MT-related metrics – not related to purpose and communication aspects – in perspective. 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 address whether the intended communication has fulfilled its purpose when this is 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 – often referred to as 'deep' NLU – will be aware of context and 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), and show consideration of the reader/listener (output/consumer view).

Only MT with deep NLU will, for example, be able to efficiently support a human-to-human or human-to-machine conversation that exhibits qualities like being contextualised, adaptive, personalised, and knowledge-rich. The following ingredients currently seem to emerge as important elements for next-generation MT (based on Deep NLU): 1. existing standards related to annotations; 2. the FAIR data principles as backbones of investment protection, and 'responsible MT'; 3. experts like translators, domain specialists, modellers, data scientists for curation; 4. more open, standardised, flexible and robust technologies for all dimensions of data management; and 5. large, multilingual translation models that are safe to use and can easily be adapted for resource-sparse computing environments, to specific tasks and domains, and for low-resource languages.

## 4 Summary and Conclusions

Nowadays MT is widely used by the general public, public sector and government agencies, SMEs, LSPs and many other industries. This will continue to grow, covering new application areas to support Europe's digital single market as well as DLE. Looking forward to 2030, we expect the movement towards deep NLU to enable efficient, 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. So far, the absence of a clear roadmap and support for LT at European level has led to an incohesive, fragmented European market with disparate language support for the language communities of Europe. We hope that the ELE SRIA (Chapter 45) will have positive effects in this regard.

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 global companies. 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. The future publicly available MT systems should not depend on non-European multinationals.

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 well-resourced languages, limiting diversity and reinforcing existing disparities. Furthermore, explainable and interpretable machine learning is attracting more and more attention, and a fundamental breakthrough is needed in the understanding of how current MT algorithms work.

The increasing quality of MT and the expanding preference for voice-based interaction points to applications for speech-to-speech translation and multimodal MT in order to break the language barrier for human communication.

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. Collection of usable language data is particularly important. If lawmakers could agree that using aligned translations of copyrighted data constitutes fair use, LT stakeholders could immediately avail of this high-quality language data. There is also a disparity between publicly available and proprietary bilingual data. A crucial breakthrough could be achieved if policy frameworks make it mandatory for Member States to make all data in natural language-related workflows publicly available.

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.) compared to large US and Chinese IT corporations. There is also an uneven distribution of resources across countries, regions and languages.

Finally, the hardware on which MT runs must be scaled down. 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. The EU has the opportunity to be a pioneer in green LT by developing efficient models and hardware.

At the level of policies/instruments, much more synchronisation of activities between national and international bodies is necessary. A desirable approach for the 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

Artetxe, Mikel, Gorka Labaka, and Eneko Agirre (2018). "Unsupervised Statistical Machine Translation". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language*

*Processing, Brussels, Belgium, October 31 - November 4, 2018*. Ed. by Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii. Association for Computational Linguistics, pp. 3632–3642. https://doi.org/10.18653/v1/d18-1399.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio (2015). "Neural Machine Translation by Jointly Learning to Align and Translate". In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. http://arxiv.org/abs/1409.0473.

Barrault, Loïc, Ondrej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri (2019). "Findings of the 2019 Conference on Machine Translation (WMT19)". In: *Proceedings of the Fourth Conference on Machine Translation, WMT 2019, Florence, Italy, August 1-2, 2019 - Volume 2: Shared Task Papers, Day 1*. Ed. by Ondrej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno-Yepes, Philipp Koehn, André Martins, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana L. Neves, Matt Post, Marco Turchi, and Karin Verspoor. Association for Computational Linguistics, pp. 1–61. https://doi.org/10.18653/v1/w19-5301.

Bērziņš, Aivars, 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, and Andy Way (2022). *Deliverable D2.13 Technology Deep Dive – Machine Translation*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/MT-deep-dive.pdf.

Bié, Laurent, 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 (2020). "Neural Translation for the European Union (NTEU) Project". In: *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. Lisboa, Portugal: European Association for Machine Translation, pp. 477–478. https://aclanthology.org/2020.eamt-1.60.

Castilho, Sheila (2021). "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)*. Online: Association for Computational Linguistics, pp. 34–45. https://www.aclweb.org/anthology/2021.humeval-1.4.

Castilho, Sheila, 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 (2019). "Large-scale Machine Translation Evaluation of the iADAATPA Project". In: *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*. Dublin, Ireland: European Association for Machine Translation, pp. 179–185.

Dou, Zi-Yi, Antonios Anastasopoulos, and Graham Neubig (2020). "Dynamic Data Selection and Weighting for Iterative Back-Translation". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Association for Computational Linguistics, pp. 5894–5904. https://doi.org/10.18653/v1/2020.emnlp-main.475.

European Commission (2018). *Coordinated Plan on Artificial Intelligence*. COM(2018) 795 final. https://digital-strategy.ec.europa.eu/en/policies/plan-ai.

European Commission (2021). *Horizon Europe Work Programme 2021-2022*. European Commission Decision C(2021)4200 of 15 June 2021. https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/how-to-participate/reference-documents.

European Union (2018a). *Proposal for a Directive of the European Parliament and of the Council on the re-use of public sector information (recast), COM(2018) 234 final*.

European Union (2018b). *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*.

Fomicheva, Marina, Piyawat Lertvittayakumjorn, Wei Zhao, Steffen Eger, and Yang Gao (2021). "The Eval4NLP Shared Task on Explainable Quality Estimation: Overview and Results". In: *Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems*. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 165–178.

Gangi, Mattia Antonino Di, Matteo Negri, Roldano Cattoni, Roberto Dessì, and Marco Turchi (2019). "Enhancing Transformer for End-to-end Speech-to-Text Translation". In: *Proceedings of Machine Translation Summit XVII Volume 1: Research Track, MTSummit 2019, Dublin, Ireland, August 19-23, 2019*. Ed. by Mikel L. Forcada, Andy Way, Barry Haddow, and Rico Sennrich. European Association for Machine Translation, pp. 21–31. https://aclanthology.org/W19-6603/.

Haddow, Barry, Alexandra Birch, and Kenneth Heafield (2021). "Machine Translation in Healthcare". In: *The Routledge Handbook of Translation and Health*. Routledge, pp. 108–129.

Han, Jesse Michael, Igor Babuschkin, Harrison Edwards, Arvind Neelakantan, Tao Xu, Stanislas Polu, Alex Ray, Pranav Shyam, Aditya Ramesh, Alec Radford, and Ilya Sutskever (2021). "Unsupervised Neural Machine Translation with Generative Language Models Only". In: *CoRR* abs/2110.05448. https://arxiv.org/abs/2110.05448.

Kenny, Dorothy, ed. (2022). *MultiTraiNMT: Machine Translation for Multilingual Citizens*. In preparation. Berlin: Language Science Press.

Koehn, Philipp (2005). "Europarl: A Parallel Corpus for Statistical Machine Translation". In: *Proceedings of Machine Translation Summit X: Papers, MTSummit 2005, Phuket, Thailand, September 13-15, 2005*, pp. 79–86. https://aclanthology.org/2005.mtsummit-papers.11.

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

Libovický, Jindřich, Helmut Schmid, and Alexander Fraser (2022). "Why don't people use character-level machine translation?" In: *Findings of the Association for Computational Linguistics: ACL 2022*. Dublin, Ireland: Association for Computational Linguistics, pp. 2470–2485.

Lopes, António V., M. Amin Farajian, Rachel Bawden, Michael Zhang, and André F. T. Martins (2020). "Document-level Neural MT: A Systematic Comparison". In: *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation, EAMT 2020, Lisboa, Portugal, November 3-5, 2020*. Ed. by Mikel L. Forcada, André Martins, Helena Moniz, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof Arenas, Mary Nurminen, Lena Marg, Sara Fumega, Bruno Martins, Fernando Batista, Luísa Coheur, Carla Parra Escartín, and Isabel Trancoso. European Association for Machine Translation, pp. 225–234. https://aclanthology.org/2020.eamt-1.24/.

Ma, Shuming, Li Dong, Shaohan Huang, Dongdong Zhang, Alexandre Muzio, Saksham Singhal, Hany Hassan Awadalla, Xia Song, and Furu Wei (2021). "DeltaLM: Encoder-Decoder Pretraining for Language Generation and Translation by Augmenting Pretrained Multilingual Encoders". In: *CoRR* abs/2106.13736. https://arxiv.org/abs/2106.13736.

Martindale, Marianna, Kevin Duh, and Marine Carpuat (2021). "Machine Translation Believability". In: *Proceedings of the First Workshop on Bridging Human – Computer Interaction and Natural Language Processing*. Online: Association for Computational Linguistics, pp. 88–95. https://aclanthology.org/2021.hcinlp-1.14.

Mason, Sarah Bawa (2019). "Joss Moorkens, Sheila Castilho, Federico Gaspari, Stephen Doherty (eds): Translation quality assessment: from principles to practice – Machine Translation: Technologies and Applications, Volume 1, Springer International Publishing, Heidelberg & Berlin, xii + 287 pp, ISBN 978-3-319-91240-0 (hardcover), 978-3-030-08206-2 (paperback), 978-3-319-91241-7 (eBook)". In: *Machine Translation* 33.3, pp. 269–277. https://doi.org/10.1007/s10590-019-09241-w.

Metuzale, Kristine, Alexandra Soska, and Marcis Pinnis (2020). "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, AMTA 2020 - Volume 2: User Papers, Virtual, October, 2020*. Ed. by Janice Campbell, Dmitriy Genzel, Ben Huyck, and Patricia

O'Neill-Brown. Association for Machine Translation in the Americas, pp. 525–546. https://aclanthology.org/2020.amta-user.25/.

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

O'Brien, Sharon, Patrick Cadwell, and Alicia Zajdel (2021). *Communicating COVID-19: Translation and Trust in Ireland's Response to the Pandemic*. Tech. rep. School of Applied Language and Intercultural Studies, Dublin City University. https://www.dcu.ie/sites/default/files/inline-files/covid_report_compressed.pdf.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). "Bleu: a Method for Automatic Evaluation of Machine Translation". In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*. ACL, pp. 311–318. https://aclanthology.org/P02-1040/.

Pinnis, Mārcis, Stephan Busemann, Arturs Vasilevskis, and Josef van Genabith (2021). "The German EU Council Presidency Translator". In: *KI – Künstliche Intelligenz*.

Popel, Martin (2018). "Machine Translation Using Syntactic Analysis". PhD thesis. Praha, Czechia: MFF UK.

Popović, Maja (2020). "Relations between comprehensibility and adequacy errors in machine translation output". In: *Proceedings of the 24th Conference on Computational Natural Language Learning*. Online: Association for Computational Linguistics, pp. 256–264.

Rehm, Georg, Katrin Marheinecke, Rémi Calizzano, and Penny Labropoulou (2023). "Language Technology Companies, Research Organisations and Projects". In: *European Language Grid: A Language Technology Platform for Multilingual Europe*. Ed. by Georg Rehm. Cognitive Technologies. Cham, Switzerland: Springer, pp. 171–185.

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

Sennrich, Rico, Barry Haddow, and Alexandra Birch (2016a). "Improving Neural Machine Translation Models with Monolingual Data". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics. https://doi.org/10.18653/v1/p16-1009.

Sennrich, Rico, Barry Haddow, and Alexandra Birch (2016b). "Neural Machine Translation of Rare Words with Subword Units". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics. https://doi.org/10.18653/v1/p16-1162.

Skadins, Raivis, Marcis Pinnis, Arturs Vasilevskis, Andrejs Vasiljevs, Valters Sics, Roberts Rozis, and Andis Lagzdins (2020). "Language Technology Platform for Public Administration". In: *Human Language Technologies – The Baltic Perspective*. Ed. by Utka Andrius, Vaicenoniene Jurgita, Kovalevskaite Jolantai, and Kalinauskaite Danguole. Vol. 328. FAIA. IOS Press, pp. 182–190.

Strubell, Emma, Ananya Ganesh, and Andrew McCallum (2019). "Energy and Policy Considerations for Deep Learning in NLP". In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. Ed. by Anna Korhonen, David R. Traum, and Lluís Màrquez. Association for Computational Linguistics, pp. 3645–3650. https://doi.org/10.18653/v1/p19-1355.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le (2014). "Sequence to sequence learning with neural networks". In: *Advances in neural information processing systems*, pp. 3104–3112.

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

Vanmassenhove, Eva, Dimitar Sht. Shterionov, and Andy Way (2019). "Lost in Translation: Loss and Decay of Linguistic Richness in Machine Translation". In: *Proceedings of Machine Trans-*

*lation Summit XVII Volume 1: Research Track, MTSummit 2019, Dublin, Ireland, August 19-23, 2019*. Ed. by Mikel L. Forcada, Andy Way, Barry Haddow, and Rico Sennrich. European Association for Machine Translation, pp. 222–232. https://aclanthology.org/W19-6622/.

Vasiljevs, Andrejs, Khalid Choukri, Luc Meertens, and Stefania Aguzzi (2019). *Final study report on CEF Automated Translation value proposition in the context of the European LT market/e-cosystem*. DOI: 10.2759/142151. https://op.europa.eu/de/publication-detail/-/publication/8494e56d-ef0b-11e9-a32c-01aa75ed71a1/language-en.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is all you need". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6000–6010.

Way, Andy (2013). "Traditional and Emerging Use-Cases for Machine Translation". In: *Proceedings of Translating and the Computer*. Vol. 35. London.

Yang, Jian, Shuming Ma, Haoyang Huang, Dongdong Zhang, Li Dong, Shaohan Huang, Alexandre Muzio, Saksham Singhal, Hany Hassan, Xia Song, and Furu Wei (2021). "Multilingual Machine Translation Systems from Microsoft for WMT21 Shared Task". In: *Proceedings of the Sixth Conference on Machine Translation, WMT@EMNLP 2021, Online Event, November 10-11, 2021*. Ed. by Loïc Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno-Yepes, Philipp Koehn, Tom Kocmi, André Martins, Makoto Morishita, and Christof Monz. Association for Computational Linguistics, pp. 446–455. https://aclanthology.org/2021.wmt-1.54.

Yao, Shaowei and Xiaojun Wan (2020). "Multimodal Transformer for Multimodal Machine Translation". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4346–4350. https://aclanthology.org/2020.acl-main.400.

Zhang, Biao, Philip Williams, Ivan Titov, and Rico Sennrich (2020). "Improving Massively Multilingual Neural Machine Translation and Zero-Shot Translation". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*. Ed. by Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault. Association for Computational Linguistics, pp. 1628–1639. https://doi.org/10.18653/v1/2020.acl-main.148.