



Neural machine translation for Hungarian

LÁSZLÓ JÁNOS LAKI*  and ZIJIAN GYŐZŐ YANG 

Hungarian Research Centre for Linguistics, Hungary

Received: April 21, 2022 • Accepted: October 1, 2022

Published online: November 30, 2022

© 2022 The Author(s)



ABSTRACT

In the scope of this research, we aim to give an overview of the currently existing solutions for machine translation and we assess their performance on the English-Hungarian language pair. Hungarian is considered to be a challenging language for machine translation because it has a highly different grammatical structure and word ordering compared to English. We probed various machine translation systems from both academic and industrial applications. One key highlight of our work is that our models (Marian NMT, BART) performed significantly better than the solutions offered by most of the market-leader multinational companies. Finally, we fine-tuned different pre-finetuned models (mT5, mBART, M2M100) for English-Hungarian translation, which achieved state-of-the-art results in our test corpora.

KEYWORDS

neural machine translation, Marian NMT, BART, mBART, mT5, M2M100

1. INTRODUCTION

Nowadays, machine translation constitutes a part of our everyday life. Recently, neural network-based solutions, especially transformer models, reached the highest performance in the area of various natural language processing tasks. The neural network-based machine translation provides significantly better quality translated texts compared to formerly existing technologies, which opens up a way to use them as a pre-translation, which could increase the effectiveness of a human translator. International publications tend to center around the English language trying to achieve the best possible translation outcomes with different model architectures. Our research has Hungarian language in its focus and we assess the performance of the different

* Corresponding author. E-mail: laki.laszlo@nytud.hu

models on Hungarian as a target language. Our research question is how these different models differ from each other in terms of translation to Hungarian and we use quantitative methods to identify distinctive features between these models. All our models and corresponding scripts presented in the scope of this publication can be found on our Github and Hugging Face websites.

2. SHORT HISTORY OF MACHINE TRANSLATION

The science of machine translation is as old as the appearance of the first computers, and is still one of the most researched areas of computational linguistics. One of the very first translation systems is the electromechanical system created by Alan Turing and his team, with the help of which it was possible to crack the most advanced encryption algorithm of the time, the so-called Enigma developed and used by the Germans during World War II (see Figure 1 (Maučec and Donaj, 2019)).

In the 1970s and 1980s, advances in computer technology made it possible to create more serious program codes. This is when Rule-Based Machine Translation (RBMT) systems emerged in the field of machine translation. Their basic idea is to use the most information possible from the text to be translated (e.g. syntactic or semantic information). The simplest early implementations were the so-called direct translation systems. The method consists of translating the text to be translated word by word based on a dictionary and then sorting it into the correct order. The advantage is that it is relatively easy to implement. However, it has the disadvantage of not being able to handle complex grammatical structures and as a consequence, it achieves poor translation quality. Later, more sophisticated systems that use parsing to produce an intermediate representation of the text to be translated, which is then transformed into an abstract target language representation using pre-defined translation rules. Finally, the target language word forms are generated from this representation. These systems can be classified according to the depth of parsing and generation, and the location of the transfer. A rule-based machine translation system with precisely written rules can produce highly accurate translations, but the generation of translation rules requires a high-quality syntactic and/or semantic parser, which is available for very few languages. Furthermore, since these rules are language-specific, they have to be defined separately for each language pair, which makes it difficult to extend the system with new languages.

In the 1990s, the advent of the internet made large amounts of digitised texts available to researchers. Further developments in computing technology have enabled our computers to

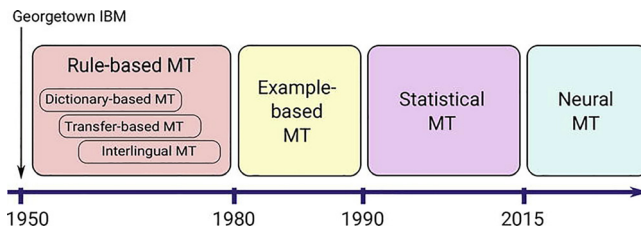


Figure 1. The different types of machine translators (Maučec & Donaj 2019)



perform complex statistical calculations on these documents in real time. This was exploited by the Statistical Machine Translation (SMT) system, which was the leading technology in the field of machine translation until the mid-2010s. The method consists of using parallel bilingual training material to learn the term transition probabilities required for translation in a supervised manner. The advantage of this method is that it offers a language-independent and easy-to-implement solution, which, depending on the size of the training material, can outperform rule-based systems.

In the 1990s, researchers tried to exploit the potential of neural networks in the field of machine translation. As early as 1997, researchers created neural network-based translation (NMT) systems with translation quality approaching that of existing dominant systems. However, since the resources available at the time were not yet suitable for handling large data sets, none of these models could be trained adequately to achieve significant results. Around the turn of the millennium, the development and proliferation of GPU technology gave researchers the opportunity to physically implement solutions that had previously only existed on paper. NMT technology has taken the lead in machine translation in a short time since 2015. Unlike SMT, the neural architecture does not work with surface word forms, but with so-called word embedding vector representations. The essence of vector representation is to have representations of words with the same meaning close to each other, while those with different meanings are far apart. This allows the system to have some world knowledge representation instead of the character form of the words.

The examples in [Figure 2](#) show¹ an example of a few words embedding representations. Notice that words used in similar contexts are plotted next to their direct synonyms.

3. TRANSFORMER-BASED NEURAL MACHINE TRANSLATION ARCHITECTURE

Over the course of the past decade, more and more architectures of NMT systems have been established, and the transformer architecture, introduced by Google in 2017, has proven to be the most successful and is still the market leader in practically all areas of language technology. The core of the model is the encoder-decoder architecture (see [Figure 3](#) (Vaswani et al., 2017)).

The encoder part is responsible for producing the linguistic representation of the source language model. The transformer architecture has the advantage of being able to process several words (100, 512, 1,024 or 2,048) at the same time, so that it can take into account not only the words of a given sentence, but also larger contexts. The output of the encoder is essentially a vector representation of the sentence to be translated. The second component is the decoder, which is responsible for generating the words of the target sentence. To do this, it takes into account the representation of the source language sentence (which is the output of the encoder) as well as the words that are already generated. The end-of-sentence signal indicates to the generator that it has finished the translation. To teach the model, as in the SMT system, only a training set of concurrently translated sentences is required, but at least an order of magnitude more material than the set used there is needed.

¹<https://medium.com/@hari4om/word-embedding-d816f643140> (Last seen: 07/09/2022)



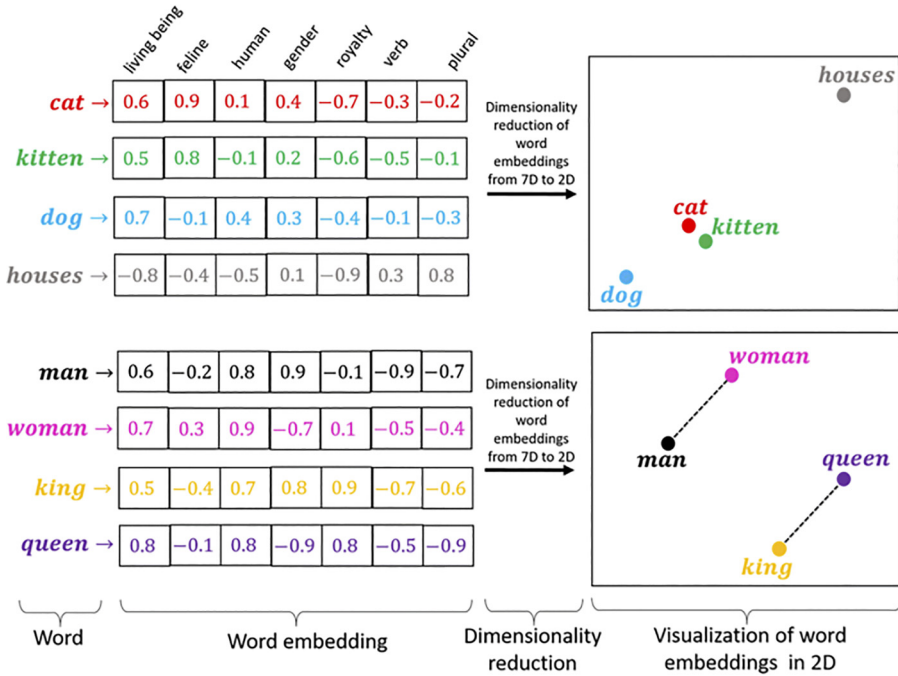


Figure 2. Example of word embedding representation

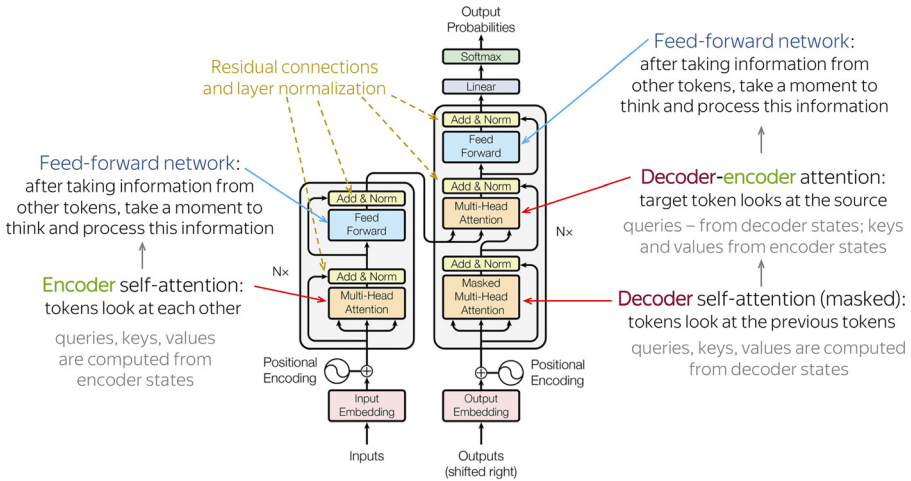


Figure 3. Transformers architecture (Vaswani et al. 2017)



The biggest advantage of NMT over SMT systems is that the translation produced by NMT is much more fluent for the human reader. Facilitated by the word embedding model, a much better translation can be achieved for agglutinating languages (such as Hungarian) since the different conjugated forms of words are not competing translations, but have become coexisting entities in the vector space thanks to the representation. Altogether this resulted in an increased willingness by human translators to use machine pre-translated text in their work.

4. AVAILABLE PRE-TRAINED MACHINE TRANSLATION MODELS AND FRAMEWORKS

In this article we provide an overview of the currently existing machine translation systems with academic or industrial purposes. As for the industrial solutions, it was not always possible to decipher the exact underlying architecture and parameters of the model. In this chapter all the systems in question are presented in detail. The following descriptive part can be considered as an in-depth literature review, since it covers all machine translation systems related to Hungarian.

4.1. Machine translation systems with research focus

Marian NMT: The Marian NMT (Junczys-Dowmunt et al. 2018) is a framework that is written in C language, which is an easy-to-install, well-documented, memory- and resource-optimized implementation. Due to these previously described characteristics, Marian NMT is the most commonly used machine translation tool by academic users and developers (Barrault et al. 2019). Marian NMT is based on an attention model supported by an encoder-decoder architecture. It is based on a neural machine translation model. Its main advantage over other methods is that it uses pre-trained language models. It can reach the fastest runtime learning without the use of pre-training language models. Transformers in two different sizes can be trained:

- Marian base (transformer-base task): 6 encoder layers and 6 decoder layers; 8 heads of attention; words embedding dimension: 512; input length: 512; pre-attached mesh size: 2,048
- Marian large (transformer-big task): 6 encoder layers and 6 decoder layers; 16 heads of attention; words embedding dimension: 1,024; 1,024 input length: 1,024; pre-attached mesh size: 4,096

BART, mBART: BART (Lewis et al. 2020) is a Transformer-based denoising autoencoder that can be used for the pre-training of sequence-to-sequence models. It has an encoder-decoder architecture. It uses a ‘noised’ source text as input, then it reconstructs the original text by predicting the corrupted parts. BART has a similar setup to BERT (Devlin et al. 2019), however, with characteristic differences in its architecture. Notably, one such difference is the additional cross-attention of the layers over the final hidden layer, which is present in BART, but not in BERT. Moreover, BART lacks the extra feed-forward network that can be found in BERT before word-prediction. The application of BART offers a high degree of flexibility regarding the usage of noising schemes, which is illustrated by the fact that any type of document corruption is compatible with the system as opposed to other denoising autoencoders. The BART practically



combines a BERT type model with a GPT type model. The difference from BERT model is that the denoising tasks are different:

- Token Masking: random tokens are sampled and replaced with [MASK] elements.
- Token Deletion: random tokens are deleted from the input.
- Text Infilling: a number of text spans (Joshi et al. 2020) are sampled, with span lengths drawn from a Poisson distribution. Each span is replaced with a single [MASK] token.
- Sentence Permutation: a document is divided into sentences based on full stops, and these sentences are shuffled in a random order.
- Document Rotation: a token is chosen uniformly at random, and the document is rotated so that it begins with that token.

Additionally, BART can be fine-tuned to be optimal for various downstream tasks, including sequence and token classification, sequence generation and machine translation. Experimentation revealed that BART performance can be especially high in large-scale pre-training, for instance, in discriminative tasks like SQuAD (Rajpurkar et al. 2016) and GLUE (Wang et al. 2018) it is comparable with RoBERTa. Furthermore, BART outperforms all previously established models in summarization tasks. Accumulating evidence suggests that BART performs the best when applied for Natural Language Generation (NLG), but achieves remarkable results in translation and comprehension tasks as well. In the recent research, we did experiments with BART base models. Two different BART models were trained for English by the Meta Research:

- BART base: 12 layer; hidden layer size: 768; 139M parameters.
- BART large: 24 layer; hidden layer size: 1,024; 406M parameters.

In our research we did experiments with our own pre-trained BART base model.

The mBART (Liu et al. 2020) is a denoising autoencoder model pre-trained on multiple languages and it is based on the seq2seq concept. It is usually applied to improve the performance of both supervised and unsupervised machine learning. The mBART follows the BART scheme in its architecture. The authors of the model put an emphasis on multilinguality during pre-training, then the model was fine-tuned for a bilingual setting. As for the pre-training, the CC25 (Wenzek et al. 2020; Conneau et al. 2020) corpus was used, which contains 25 languages and the texts are extracted from the CommonCrawl database (Kúdela et al. 2017). The multilingually pre-trained model was used for both sentence- and document-level machine translation. It is particularly important to point out that only with the application of the seq2seq concept could improve the quality of document-level machine translation, this is a significant step forward compared to previous research work (Miculicich et al. 2018; Li et al. 2019). The experimental results with the mBART model highlight the true potential of multilingual pre-training with an applicability for transfer learning.

The mBART model has an extended version, which is called mBART-50 (Tang et al. 2020). Using the original mBART model extra 25 languages were added to support multilingual machine translation models of 50 languages.

mBART models, even the mBART-50 model, do not include Hungarian language capabilities, but taken the advantage of sentencepiece tokenization (Kudo & Richardson 2018), we could adopt this model for Hungarian. In our hypothesis, the more machine translation and different language knowledge it has, the more knowledge can be adopted for our English-Hungarian. Thus, we used the mbart-large-50-many-to-many-mmt model checkpoint



in our research. This model is a fine-tuned checkpoint of mBART-large-50 that is trained for multilingual machine translation tasks. The model can translate directly between any pair of 50 languages.

T5, mT5: The T5 (Text-To-Text Transfer Transformer) (Raffel et al. 2020) is a model and framework developed by the Google research team, which offers a new perspective to solve natural language processing tasks. Transfer learning constitutes an important part of the natural language processing toolbox. In the course of transfer learning the language model is trained on a data-rich task followed by a fine-tuning step for a subsequent specific task.

In an ideal case, the model can acquire general knowledge during the pre-training phase that is transferable and it can be further applied to the specific tasks. The T5 project applies transfer learning principles in the context of the seq2seq approach. The initial idea was that all language processing tasks (translation, question answering, classification) should be considered as a text-to-text issue, therefore the input is a text and the output will be another text. The great advantage of the text-to-text paradigm is their wide range of applicability, since it can be used for practically any natural language processing task, for example machine translation, summary generation, question answering or sentiment analysis.

Such large-scale experiments require special corpora. For this aim, the Colossal Clean Crawled Corpus (abbreviated as C4) was created, which is an English language extract of hundreds of gigabytes of the World Wide Web, collected and cleaned. The C4 corpus is based on the CommonCrawl 5 database. Another important feature of transfer learning methods is that non-labelled datasets are required for their pre-training. Additional requirements for such corpora are to be designated as large enough, diverse and high-quality. As an example, the C4 corpus used is two times the size of Wikipedia, therefore, it contains significantly more data. In the case of T5, based on the number of parameters 5 different models have been created:

- Small (300 million parameters), Base (580 million parameters), Large (1.2 billion parameters), XL (3.7 billion parameters), XXL (13 billion parameters)

As a result of the T5 project a highly efficient framework has been created, which produces excellent results. One of the models with 11 billion parameters reached outstanding performance at several benchmarks including GLUE, SuperGLUE, SQuAD and CNN/Daily Mail reference tasks.

The mT5 (Xue et al. 2021) extends the above detailed T5 to several languages. The authors attempted to preserve the structural features of T5, which previously proved to be successful in several experiments. In line with this strategy, mT5 inherited the text-to-text problem approach and the general pre-learning process with the application of large corpora.

In order to train mT5, the mC4 corpus was used. mC4 is the multilingual version of C4 with texts from 101 different languages. A common problem with multilingual models is the unequal representation of languages. If a certain language is underrepresented in the corpus, inappropriate fitting by the model can occur due to a higher sampling rate. To address this issue the authors applied a frequency-based sampling procedure, which has already been used in previous work as well (Devlin et al. 2019; Aharoni et al. 2019). Considering the fact that the mT5 model was trained on a corpus of more than 100 languages it was therefore necessary to use a larger dictionary consisting of 250,000 word pieces.

To evaluate the performance of the mT5, 6 tasks from the XTREME multilingual reference system (Hu et al. 2020) were applied. Several tasks are included in this reference framework,



e.g. sentence pairing, noun recognition and question answering. In terms of question answering tasks, mT5-XXL (largest model in the mT5 framework) achieved the highest performance. The mT5 project revealed that the framework in question can be successfully applied in a multilingual context and the models can achieve outstanding results in various reference tasks.

mT5 includes Hungarian language as well, therefore, we applied the mT5 small and mT5 base models for machine translation in our research.

M2M100: M2M100 (Aharoni et al. 2019) is a project in the Fairseq multilingual machine translation pipeline. The aim of multilingual machine translation is to create a comprehensive model that can translate from any language to any other languages. For a long time, machine translation was considered to be rather English-centric, i.e. the majority of language models have been created that translate from English to other languages and vice versa. However, translation in real life is not used in such an exclusive manner. Translation from and to many other languages other than English is required and there is great demand for these types of translation services. The M2M100 project resulted in a translation tool and dataset for 100 languages, which is a highly diversified machine translation by shifting from an English-centered approach to multilinguality and paves the way towards novel methodological breakthroughs. In the case of machine translation from multiple languages to multiple languages, the creation of large datasets is necessary. The generation of such a large volume of multilingual data requires data mining (Artetxe & Schwenk 2019) and reverse translation (Sennrich et al. 2016).

M2M100 includes Hungarian language knowledge, therefore, we also applied it in our research and assessed its performance. Furthermore, we further fine-tuned the model on our corpora. There are two versions of M2M100 are available: M2M100_418M (418 million parameters) and M2M100_1.2B (1.2 billion parameters).

Helsinki Marian NMT: The HNMT (Helsinki Neural Machine Translation) (Tiedemann & Thottingal 2020) is based on (base) Marian NMT (Junczys-Dowmunt et al. 2018), which is currently the best performing translation method from English to Finnish, moreover, it reached the highest BLEU values for this language pair. The performance of HNMT was tested on English-to-Latvian, English-to-Chinese and Chinese-to-English language pairs and directions, however, it achieved only moderate results. The HNMT machine translation system works particularly well in the case of morphologically rich languages, such as Finnish. One major goal of the team from the University of Helsinki was to create machine translation models for as many languages as possible. The group created several base models for English-Hungarian as well, which were tested in our research.

4.2. Machine translation systems developed by the industry

DeepL: DeepL Translate is a freely available online translation system (DeepL GmbH, Cologne, Germany). The company behind the translator tool is Linguee. The company launched a search engine in 2009, which specializes in translation (deepl.com). The DeepL Translate uses convolutional neural networks (Kim 2014), and thanks to its architecture it can produce more polished and naturally sounding translations compared with the solutions by other competitors on the market. Launched in 2018, DeepL Pro is a further optimized version of the company's proprietary artificial intelligence solutions and it provides even higher quality machine translation. In 2021, 13 European languages, including Hungarian, were added to the DeepL repertoire.



Google Translate: The Google Translate (Wu et al. 2016) was launched in 2003. During the first phase, it operated based on a statistical machine translation principle, which was replaced by neural network-based machine translation in 2016. The introduction of the neural network-based approach has significantly improved the quality of the translation by providing inferences on a broader context and thus more authentic translations. The Google Translate lists several types of different translated versions, for example in the case of languages with female and male distinction (e.g. French or Spanish) first the feminine and then the masculine version appears (Rescigno et al. 2020). Google Translate can handle 109 different languages. Since 2020, there has been an additional feature that enables translating spoken text as well.

Yandex: Yandex is a Russia-based technology company that provides machine translation solutions on the market of digital products. The translation system consists of two separate machine translation systems: a statistical machine translation tool, which contains hundreds of thousands of texts with the same information but written in different languages. The Yandex Statistical Translator is a three-component machine translation system: the translation model, the language model, and the decoder. The actual translation process is done by the decoder. It uses the different translated versions by the translation model, and creates a frequency-based ranking, which is then determined by the language model. The other main component is a neural machine translation system with an encoder-decoder architecture, which has specifically an RNN architecture according to the available information (Cho et al. 2014). Finally, the system compares the translation output from the two translation subsystems by the CatBoost algorithm (Prokhorenkova et al. 2018) and then outputs the best translation as the final output.

Bing Translate – Microsoft Translate: The Bing Translator constitutes a part of the Microsoft Cognitive Services product family. It is capable of translating texts into more than 100 different languages. Since 2021, it provides a solution for translating entire documents. Initially, it was based on statistical machine translation, which was replaced by a neural network-based approach in 2018. Microsoft is dedicated to the development of advanced solutions in multilingual machine translation and the company heavily invests in research to improve efficiency and accuracy. Xu Tan et al. have developed a tool (Tang et al. 2020) to overcome the difference in the accuracy between multilingual and monolingual models, which is based on the knowledge distillation principle (Bucila et al. 2006). Knowledge distillation was initially used to develop more effective models by making them ‘slimmer’. The core principle behind knowledge distillation is that there is a ‘student model’, that can achieve the performance of a ‘teacher model’ or a set of models. The implementation of this idea to machine translation means that there are teacher models that are specialized for each language pair and train the student model, which will be capable of handling all the language pairs as a result of the training. Two different procedures have been developed: one is selective distillation, where the use of distillation is based on a performance threshold, and the other is Top-K distillation, which uses the probability distribution provided by the teacher models, and only the models with the best coefficient are loaded into the memory. The effectiveness of this methodology is highlighted by its superior performance in the translation of TED talk transcripts from 44 languages to English: it could reach a BLEU-score improvement of 1 or even higher.

eTranslation: eTranslation is an automated translation tool that can be used to translate texts or entire documents into official languages of the Member States of the European Union, as well as Icelandic, Norwegian, Russian and simplified Chinese. The translation tool is provided by the European Commission with the intention to support small and medium-sized companies in the European Union, moreover to facilitate smooth and effective communication between public



service providers, administrative officials, and SMEs. The eTranslation tool can be easily integrated with other digital solutions if translation capability is required. Several processing steps and text filtering options are also available under the CEF eTranslation Building Block project to make the machine translation easier. For example, long sentences are first divided into smaller parts before translation and then reassembled to a coherent text. The eTranslation system has been trained on texts with specialized content, such as tenders, legal and medical texts, etc. The model has been trained on more than 1 billion sentences in 24 different languages.

Baidu Translate: In 2019, Baidu published a paper (Sun et al. 2019) in which state-of-the-art results were achieved in a case-sensitive Chinese-English task and second in an English-Chinese task. The Baidu translation system is based on neural transformer model, which uses a monolingual pre-trained encoder. To increase the performance of the system, a deeper and bigger transformer was used. For the better representation of the source sentences, the number of encoder layers was increased (from 6 to 30 for the base version and from 6 to 15 for the big version) and the dimension of feed-forward network was increased from 4,096 to 15,000 for the big version. During the pre-training process, reverse-translation mechanism, joint training, data augmentation, and knowledge distillation approaches were applied. To gain the best output performance, model ensemble and re-ranking techniques were integrated into the architecture. The online application supports 201 different languages.

4.3. Training and test corpora

We generated our own English-Hungarian parallel corpus for machine translation purposes. To build the corpus, we took English-Hungarian (en-hu) parallel sub-corpora from the OPUS corpus (Tiedemann 2012) and the Hunglish corpus Varga et al. (2007). Here we list the subcorpora of OPUS that we used to generate our corpus: OpenSubtitles, Tatoeba, WikiMatrix, EUbookshop, PHP manual, TED2020, KDEdoc, KDE4. The sizes of the corpora are shown in Table 1 (on not tokenized text).

For testing our models, we used two corpora. The first one is from our OPUS corpus. We chose 10,000 randomly selected segments that our training corpus does not contain. The second test corpus is the official devtest subcorpus of Hunglish from Shared Task of WMT 2009 (WMT09).²

4.4. Proprietary trained machine translation models

During our research, two Marian, one BART, one MBart and two mT5 machine translation models were trained for the English-Hungarian language pair. To date, we pioneered to train an English-Hungarian bilingual BART base model for the first time. The model is available on our Hugging Face site. For the training, the WikiText-103 (Merity et al. 2017) and Hungarian Wikipedia part of Webcorpus 2.0 (Nemeskey 2020) were used. In the original setting, only those paragraphs were included that contained at least one punctuation mark. The dimensions of the corpora are shown in Table 2 (on tokenized text).

For the pre-training of our English-Hungarian BART-base model, we used 4 pieces of GeForce GTX 1080 Ti (12 GB) video cards with the following parameters: batch size/GPU: 12;

²<https://www.statmt.org/wmt09/translation-task.html>



Table 1. Size of training subcopora

	Segment	Token		Type		Avg. token/ sentence	
		en	hu	en	hu	en	hu
OpenSubtitles	42,655,519	272,571,665	209,481,645	2,382,239	6,519,406	6.39	4.91
ParaCrawl	12,681,746	196,278,983	172,671,171	3,555,484	5,713,776	15.48	13.62
WikiMatrix	488,319	8,978,943	7,673,323	627,814	1,057,487	18.38	15.71
TED2020	308,341	5,194,871	3,982,056	158,210	495,452	16.85	12.91
EUbookshop	438,264	9,406,548	7,847,111	360,311	648,778	21.46	17.90
KDE4	120,657	622,959	649,457	62,257	98,940	5.16	5.38
Tatoeba	109,041	639,834	505,838	30,759	84,570	5.86	4.64
PHP	35,423	169,610	157,583	17,215	25,854	4.79	4.45
KDEdoc	861	10,904	9,474	2,402	2,987	12.66	11.00
Hunglish	1,520,610	26,784,043	21,565,337	483,581	1,192,213	17.61	14.18
SUM	58,358,781	520,658,360	424,542,995	5,309,559	8,350,079	16.22	14.62

Table 2. Corpora sizes of BART pre-training task

	English WikiText-103	Hungarian Wikipédia
Segment	707,391	1,098,156
Token	96,534,563	90,349,849
Type	596,820	3,137,980
Avg. sentence/paragraph	5	4
Avg. sentence/paragraph	125	69

dictionary size: 40,000; learning rate: $2e^{-8}$; number of learning steps: 170,000. For the pre-training the *Seq2SeqTrainer* and the *BartForCausalLM* functions were used that can be found in Hugging Face Transformers library. We further fine-tuned our BART model for machine translation with the English-Hungarian language pair. For fine-tuning, we used 4 GeForce GTX 1080 (12 GB) video cards with the following parameters: batch size/GPU: 26; maximum text length (input and output): 128; warmup: 15,000; fp16; epoch: 10; learning rate: $5e^{-5}$. The example code available in Hugging Face Transformers library was used for fine-tuning. In our machines translation fine-tuning experiments, we were working with two variants: *BART-128* and *BART-512*. The only difference is the input sequence length.

In the case of Marian NMT we applied the default parameter settings defined by the framework. First, we used a base model (Marian base), in the second case we applied twice as many parameters (Marian big). The models are available on Hugging Face. As for the subword



tokenization the built-in Sentence Piece (Kudo & Richardson 2018) tokenizer was used. Size of the dictionary: 32,000.

In our experiment, with M2M100, mBART and mT5 we fine-tuned a pre-trained M2M100, mbart-large-50-many-to-many-mmt (mBART-mmt), a mT5 small and a mT5 base model to English-Hungarian translation:

- **mBART-mmt**: In our experiment, we fine-tuned the facebook/mbart-large-50-many-to-many-mmt (Tang et al. 2021) model. This model is a fine-tuned mBART-large-50 multilingual Sequence-to-Sequence model, which is created using the original mBART model and extended to add extra 25 languages to support multilingual machine translation models of 50 languages. The 50 languages do not contain Hungarian, thus we added the “hu_HU” language code as a special token to the vocabulary. This experiment shows that despite the fact that there was no Hungarian language knowledge in the model, it is still can be fine-tuned for Hungarian, since the 50 languages could contain Hungarian text fragments. For machine translation, the MBartTokenizer and MBartForConditionalGeneration from the Hugging Face Transformers library were used.
- **M2M100**: Two models are available, a smaller (418M) and a larger (1.2B). For machine translation, the M2M100Tokenizer and M2M100ForConditionalGeneration from the Hugging Face Transformers library were used. In our research we tested the models on our Hungarian corpus, then we further fine-tuned the large (1.2B) model on it.
- **mT5**: We have fine-tuned a mT5 small and a mT5 base model for English-Hungarian machine translation. For fine-tuning, the MT5Tokenizer and MT5ForConditionalGeneration from the Hugging Face Transformers library were used.

To train the mT5 small model we used $4 \times$ GeForce GTX 1080 (12 GB) video card with the following parameters: batch size: 6; prefix: “translate English to Hungarian: ”, maximum text length (input and output): 128; epoch: 1; learning rate: $5e^{-5}$. Unfortunately, the epoch number was set to 1 only, which resulted in a running time of almost a month. In the case of M2M100, mBART and mT5 base models, we used 4 x NVIDIA A100 (80 GB) video card with the following parameters: batch size: 12 (mBART), 22 (mT5); maximum text length (input and output): 256; epoch: 1; learning rate: $5e^{-5}$; prefix (mT5): “enhu: ”. To fine-tune these models the same library was used as in the case of BART fine-tuning.

In Table 3, you can see the most important technical information and training parameters of our models. Since, the detailed technical information of machine translation systems for industries is usually not published, or we cannot be sure about the architecture of the currently available version, we could not present them in our comparison. In Table 3, the first block shows the information of our custom-trained models. In the second block, you can see the information on fine-tuned models. As you can see the details, the different models were trained in different environments. In the parameters column, you can see the approximated (\sim) parameter numbers.

4.5. Applied and tested machine translator systems and models

In our research we probed a range of different methods applied for various research and industrial applications that include the capability to perform machine translation from English into Hungarian. In our experiments, we tested the following systems and models:



Table 3. Technical and training information of our models

	batch/device	epoch	time	machine	Parameters (~)
Marian big	99	23	3 weeks	4 × GTX 1080 (12 GB)	530 million
Marian base	146	15	5.5 day	4 × GTX 1080 (12 GB)	240 million
BART-512	4	1	3 weeks	4 × GTX 1080 (12 GB)	139 million
BART-128	26	10	6 weeks	4 × GTX 1080 (12 GB)	139 million
M2M100_1.2B ft	16	2	1 week	4 × A100 (80 GB)	1.2 billion
mBART-mmt ft	24	2	1 week	4 × A100 (80 GB)	680 million
mT5 base ft	22	1	4 days	4 × A100 (80 GB)	580 million
mT5 small ft	5	1	4 weeks	4 × GTX 1080 (12 GB)	300 million

- **Helsinki Marian NMT:** several of their models include English-Hungarian translation capabilities: English-Hungarian (en-hu); English-Finnugorese (en-fiu); English-Urish (en-urj); English-multi (310 languages) (en-multi). For translation, we used the MarianTokenizer and MarianMTModel from the Hugging face Transformers library.
- **eTranslation:** it provides free service opportunity for academic scholars. Followed registration, we submitted our test file. The translated version was sent by e-mail.
- **deepL:** We translated using the online file translation function. Each file contained 500 sentences.
- **Google:** Translated using the online website translation function. We inserted the source sentences into a website.
- **Microsoft:** We translated using Azure Translator, a cloud-based document translator module of the Microsoft Bing services.
- **Yandex:** Translated using the online document translation function. Each document contained 500 sentences.
- **Baidu:** Translated using the online website translation function. We inserted the source sentences into a website, each site contained 1,500 sentences.

5. RESULTS

The SacreBLEU (Papineni et al. 2002) (Post 2018) and chrF (Popović 2015) metrics were used to evaluate the different models and systems. In addition to the BLEU metrics, we chose the chrF metrics due to its character-based feature, which results in a more accurate evaluation in the case of agglutinating languages such as Hungarian. For the chrF evaluation, the default 6-gram character precision and the 3-gram precision was calculated as well. We have chosen this metric, because in this case a successful translation of the stem will be taken into account even if its suffix is not translated correctly.

Table 5 shows the results in the case of the different machine translators. First of all, the bottom section contains the result of the commercial solutions. We know that these systems are constantly evolving, therefore, we present our results that reflect the performance state of 2022 August.



Among the industrial applications, in our test corpora, Baidu can be seen as a winning solution. After that, eTranslation and deepL performed the best and the difference between them is not statistically significant. The systems developed by large companies (Google, Microsoft) fall into the second quality category, while Yandex is far behind. In the Hunglish corpus, the ranking is different. The winner is the deepL, then Microsoft and Google achieved the highest performance. After them, Baidu and eTranslation followed, and Yandex finished in last place here as well.

In the next section of the table above the open-source solutions can be seen. Out of the models used in our research is clearly the Helsinki en-hu model, which is not surprising, given the high level of overlap in the used training material, as well as the fact that the model is bilingual and not multilingual. On the other hand, the multi-lingual decoder Helsinki model has the worst result in our comparison. This is an interesting experiment of ours, because we intend to do research in the field of multi-lingual models.

The considerably larger pre-trained M2M100 model – despite being capable of translation to 100 languages – finished slightly behind the commercial solutions. The biggest difficulty with the model is that it requires huge technical resources to be able to fine-tune it. An 8 GPU server with video card drivers (Nvidia GTX 1080 Ti) is not enough to start a single fine-tuning on it.

The second section of the table contains our fine-tuned pre-trained language model-based translators. As expected, most of our models significantly outperform the previous ones from the sections below them. The common feature of these trainings is that these require special GPUs for deep learning, which have much larger memory capacity. Secondly, the models, which were basically trained as a translator (mBART-mmt, M2M100) outperform the ones, in which

Table 4. Comparison examples of the translator outputs

Source	- Oh, no. If you think you're tucking me away somewhere, you've got another think coming.
Reference	Ha azt tervezi, hogy bedug valahová, akkor terveljen ki valami mást.
Google	- Óh ne. Ha azt hiszed, hogy elrejtész valahova, akkor más gondolat jön.
	- <i>Oh no. If you think you're going to hide it somewhere, you'll have another thought.</i>
M2M100 ft	Ha azt hiszed, hogy elrejtethsz valahová, akkor másképp is gondolhatod.
	<i>If you think you can hide it somewhere, you may think otherwise.</i>
mBART ft	Ha azt hiszed, hogy elrejtethsz valahol, akkor nagyon tévedsz.
	<i>If you think you can hide it somewhere, you are very wrong.</i>
mT5 base ft	Ha azt hiszed, hogy elrángatsz valahol, akkor jön egy másik gondolat.
	<i>If you think you pulls me away somewhere, another thought will come.</i>
BART	Ha azt hiszed, hogy el akarsz dugni valahova, akkor másra is gondolhatsz.
	<i>If you think you want to hide somewhere, you can think of something else.</i>
Marian big	Ha azt hiszed, hogy eldughatsz valahova, akkor tévedsz.
	<i>If you think you can hide somewhere, you are wrong.</i>



pre-training was only a common task (BART, mT5). Our mT5 models were only fine-tuned at 1 epoch due to the lack of available resources, so these could not achieve as high results as expected. However, it still performed competitively, outperforming most of the industrial and research applications, despite the fact that the mT5 was originally trained for mixed tasks and the training material was only presented to it once. In our test corpora, M2M100_1.2B and mBART-mmt could achieve state-of-the-art results among all systems, but in the case of Hunglish corpus, these models could not outperform the deepL solution listed amongst the industrial systems. Taken together, all these solutions achieved higher results than most of all other (not proprietary) industrial and academic models.

Finally, in the top section of the results, our “from the scratch trained” models can be seen. The Marian big and BART-128 models were in the best quality class with almost identical results. As expected, the Marian big model achieved the best performance among our own

Table 5. Results of the different MT systems and models. The best results are highlighted with bold characters.

	OPUS			Hunglish (WMT09)		
	BLEU	chrF-3	chrF-6	BLEU	chrF-3	chrF-6
Marian big	35.54	60.12	55.64	25.09	56.81	51.33
Marian base	33.03	58.32	53.58	19.49	53.05	46.97
BART-128	34.73	59.16	54.67	21.13	54.30	48.49
BART-512	33.11	58.01	53.30	17.75	51.64	45.45
M2M100_1.2B ft	37.84	61.47	57.31	21.64	54.97	49.13
mBART-mmt ft	37.44	61.07	56.82	22.38	55.33	49.51
mT5 base ft	33.33	58.22	53.57	11.64	49.24	42.84
mT5 small ft	27.69	53.73	48.57	7.34	45.60	38.62
Helsinki en-hu	27.21	55.03	49.82	18.08	52.34	46.06
Helsinki en-fiu	24.23	52.68	47.16	15.46	49.96	43.39
Helsinki en-urj	24.16	52.56	47.09	15.60	50.02	43.45
Helsinki en-multi	14.39	43.69	36.74	8.30	41.72	33.93
M2M100_1.2B	21.62	50.93	45.73	17.76	51.88	45.73
M2M100_418M	18.75	48.40	42.72	15.22	49.50	42.88
Baidu	30.57	57.25	52.60	18.99	53.04	47.06
eTranslation	28.29	56.00	51.27	18.37	52.71	46.77
deepL	26.54	56.06	51.01	22.90	56.40	50.70
Google	25.30	54.09	49.06	20.29	53.62	47.81
Microsoft	25.22	53.02	48.00	20.47	54.75	48.70
Yandex	19.22	49.78	43.94	8.66	45.84	38.84



models, due to the fact that it works with a network comprising the highest number of parameters. In the competition of the fine-tuned models, on OPUS corpus, the fine-tuned M2M100 could gain the highest results (among all models as well). But on Hunglish corpus, our Marian big could achieve state-of-the-art results.

The most prominent advantage of the Marian model is that it can be used without GPU, therefore it needs less resources and technical requirements. But since this model was trained from scratch, that is why it needed much more epoch and time during the training phase. In contrast, the pre-trained mBART or M2M100 needed only 2 epoch to achieve higher results than our Marian big model, but these models need high-performance GPUs for fine-tuning and also for the translation generation. On Hunglish corpus the Marian big model has higher

Table 6. Second comparison examples of the translator outputs

Source	This may not make much sense to you, sir, but I'd like to ask your permission to date your daughter.
Reference	Szeretném megragadni az alkalmat uram, hogy az engedélyét kérjem, hogy találkozassak a lányával.
Google	Lehet, hogy ennek nem sok értelme van, uram, de szeretném engedélyét kérni a lányával való randevúzáshoz.
	<i>(This may not make much sense, sir, but I would like to ask your permission to date your daughter.)</i>
M2M100	Talán nem sok értelme van, uram, de szeretném az engedélyét kérni, hogy randizhassak a lányával.
	<i>It may not make much sense, sir, but I would like to ask your permission to date your daughter.</i>
mBART	Lehet, hogy önnek nem sok értelme van, uram, de szeretnék engedélyt kérni, hogy randizhassak a lányával.
	<i>You may not be making much sense, sir, but I would like to ask permission to date your daughter.</i>
mT5	Talán nem sok értelme van, uram, de szeretném kérni az engedélyét, hogy randizzon a lányával.
	<i>It may not make much sense, sir, but I would like to ask your permission to date your daughter.</i>
BART	Lehet, hogy önnek nincs sok értelme, uram, de szeretném az engedélyét kérni, hogy randizhassak a lányával.
	<i>You may not be making much sense, sir, but I would like to ask your permission to date your daughter.</i>
Marian big	Ennek talán nincs sok értelme, uram, de szeretném az engedélyét kérni, hogy randizhassak a lányával.
	<i>This may not make much sense, sir, but I would like to ask your permission to date your daughter.</i>



performance, this may be related to the training epoch number. The fine-tuned models only “saw” the Hunglish corpus twice, while the Marian big “saw” it 23 times (since it was trained from scratch).

In the second quality category of the first block, we could find Marian-base, BART-512 models. It was an interesting observation for us that there is only a 2% performance difference between Marian-big and Marian-base models, while the other common paired models have a much wider gap. Secondly, the base model size (300 MB) is half of the big ones (750 MB) and the training took only 90 h instead of 500 h of the big model. For a commercial company this price for only 2% performance gain would be clearly a no-go.

In Tables 4 and 6 we demonstrate example sentences, in which the translations of the most interesting systems are highlighted. After having examined the translations, we could conclude that the outputs by the systems were readable texts, and that the differences between the translations were mainly grammatical structural differences. This phenomenon was also reflected in the example sentences: the main reason for the erroneous translations was usually the alterations in content due to conjugation. In the presented examples, we observed that the translations made by BART, mBART, M2M100, and Marian big models had the best capability to capture the meaning of the source sentences, despite the fact that it did not correspond to the reference sentence at the character level.

6. CONCLUSION

In our current research we trained and tested different neural machine translation models and systems for the English-Hungarian language pair. We experimented with machine translation methods and systems from both academic and industrial sources. In addition to the existing models, our own machine translation systems were trained as well. We have trained two Marian NMT systems, one base and one large model. We also trained our proprietary BART model, which was then fine-tuned for machine translation. Finally, pre-trained M2M100, mBART and mT5 models were fine-tuned for English-Hungarian machine translation. The results of custom pre-trained models demonstrated that the trained large Marian NMT model and the BART model achieved significantly higher performance compared to all other models. In a comparative assessment of the two models, BART performance was surpassed by the Marian Big model by only a minimal value, which is an interesting result, since BART was able to achieve this fairly competitive performance with fewer parameters. In the experiments of fine-tuned models, M2M100 and mBART models could achieve state-of-the-art results on the OPUS corpora. A noteworthy result is that the multilingual pre-trained models can be adapted for Hungarian, even without Hungarian knowledge, as it is in the case of mBART.

REFERENCES

- Aharoni, Roei, Melvin Johnson and Orhan Firat. 2019. Massively multilingual neural machine translation. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers). 3874–3884.



- Artetxe, Mikel and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics* 7. 597–610.
- Barrault, Loïc, Ondřej 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). *Proceedings of the Fourth Conference on Machine Translation, Vol. 2: Shared Task Papers, Day 1*. 1–61.
- Bucila, Cristian, Rich Caruana and Alexandru Niculescu-Mizil. 2006. Model compression. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '06)*. 535–541.
- Cho, Kyunghyun, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1724–1734.
- Conneau, Alexis, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 8440–8451.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Vol. 1 (Long and Short Papers)*. 4171–4186.
- Hu, Junjie, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. *Proceedings of the 37th International Conference on Machine Learning (ICML)*. 4411–4421.
- Joshi, Mandar, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics* 8. 64–77.
- Junczys-Dowmunt, Marcin, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. *Proceedings of ACL 2018, System Demonstrations*. 116–121.
- Kim, Yoon. 2014. Convolutional neural networks for sentence classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1746–1751.
- Kúdela, Jakub, Irena Holubová and Ondřej Bojar. 2017. Extracting parallel paragraphs from common crawl. *The Prague Bulletin of Mathematical Linguistics* 107(1). 39–56.
- Kudo, Taku and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 66–71.
- Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7871–7880.
- Li, Liangyou, Xin Jiang and Qun Liu. 2019. Pretrained language models for document-level neural machine translation. *arXiv. abs/1911.03110*.



- Liu, Yinhan, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics* 8(11). 726–742.
- Maučec, Mirjam Sepesy and Gregor Donaj. 2019. Machine translation and the evaluation of its quality. In A. Sadollah and T.S. Sinha (eds.) *Recent trends in computational intelligence*. Rijeka: IntechOpen. 143–162.
- Merity, Stephen, Caiming Xiong, James Bradbury and Richard Socher. 2017. Pointer sentinel mixture models. 5th International Conference on Learning Representations.
- Miculicich, Lesly, Dhananjay Ram, Nikolaos Pappas and James Henderson. 2018. Document-level neural machine translation with hierarchical attention networks. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2947–2954.
- Nemeskey, Dávid Márk. 2020. Natural language processing methods for language modeling. Doctoral dissertation. Eötvös Loránd University, Budapest.
- Papineni, Kishore, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. 311–318.
- Popović, Maja. 2015. chrF: character n-gram F-score for automatic MT evaluation. *Proceedings of the Tenth Workshop on Statistical Machine Translation*. 392–395.
- Post, Matt. 2018. A call for clarity in reporting BLEU scores. *Proceedings of the Third Conference on Machine Translation: Research Papers*. 186–191.
- Prokhorenkova, Liudmila, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush and Andrey Gulin. 2018. CatBoost: Unbiased boosting with categorical features. *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS'18)*. 6639–6649.
- Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research* 21(140). 1–67.
- Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2383–2392.
- Rescigno, Argentina Anna, Johanna Monti, Andy Way and Eva Vanmassenhove. 2020. A case study of natural gender phenomena in translation: A comparison of Google Translate, Bing Microsoft Translator and DeepL for English to Italian, French and Spanish. *Workshop on the Impact of Machine Translation (iMPacT 2020)*. 62–90.
- Sennrich, Rico, Barry Haddow and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Vol. 1: Long Papers*. 1715–1725.
- Sun, Meng, Bojian Jiang, Hao Xiong, Zhongjun He, Hua Wu and Haifeng Wang. 2019. Baidu neural machine translation systems for WMT19. *Proceedings of the Fourth Conference on Machine Translation, Vol. 2: Shared Task Papers, Day 1*. 374–381.
- Tang, Yuqing., Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *ArXiv*. abs/2008.00401.
- Tang, Yuqing, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu and Angela Fan. 2021. Multilingual translation from denoising pre-training. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 3450–3466.



- Tiedemann, Jörg and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT). 479–480.
- Tiedemann, Jörg. 2012. Parallel data, tools and interfaces in OPUS. In N.C.C. Chair, K. Choukri, T. Declerck, M.U. Dogan, B. Maegaard, J. Mariani, J. Odijk and S. Piperidis (eds.) Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). Istanbul: European Language Resources Association (ELRA). 2214–2218.
- Varga, Dániel, Péter Halácsy, András Kornai, Viktor Nagy, László Németh and Viktor Trón. 2007. Parallel corpora for medium density languages. In N. Nicolov, K. Bontcheva, G. Angelova and R. Mitkov (eds.) Recent advances in natural language processing IV: Selected papers from RANLP-05. Amsterdam: Benjamins. 247–258.
- 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 I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett (eds.) Advances in neural information processing systems, Vol. 30. Curran Associates, Inc. 5998–6008.
- Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy and Samuel Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. 353–355.
- Wenzek, Guillaume, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. Proceedings of the 12th Language Resources and Evaluation Conference. 4003–4012.
- Wu, Yonghui, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. CoRR. abs/1609.08144.
- Xue, Linting, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA: Association for Computational Linguistics. 483–498.

Open Access. This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited, a link to the CC License is provided, and changes – if any – are indicated. (SID_1)

