Annales Mathematicae et Informaticae
57 (2023) pp. 92–106
DOI: https://doi.org/10.33039/ami.2022.11.001
URL: https://ami.uni-eszterhazy.hu

Solving Hungarian natural language processing tasks with multilingual generative models

Zijian Győző Yang, László János Laki

Hungarian Research Centre for Linguistics
{yang.zijian.gyozo,laki.laszlo}@nytud.hu
MTA-PPKE Hungarian Language Technology Research Group
Pázmány Péter Catholic University,
Faculty of Information Technology and Bionics
{yang.zijian.gyozo,laki.laszlo}@itk.ppke.hu

Abstract. Generative ability is a crucial need for artificial intelligence applications, such as chatbots, virtual assistants, machine translation systems etc. In recent years, the transformer-based neural architectures gave a huge boost to generate human-like English texts. In our research we did experiments to create pre-trained generative transformer models for Hungarian language and fine-tune them for multiple types of natural language processing tasks.

In our focus, multilingual models were trained. We have pre-trained a multilingual BART, then fine-tuned it to various NLP tasks, such as text classification, abstractive summarization. In our experiments, we focused on transfer learning techniques to increase the performance. Furthermore, a M2M100 multilingual model was fine-tuned for a 12-lingual Hungarian-Centric machine translation. Last but not least, a Marian NMT based machine translation system was also built from scratch for the 12-lingual Hungarian-Centric machine translation task.

In our results, using the cross-lingual transfer method we could achieve higher performance in all of our tasks. In our machine translation experiment, using our fine-tuned M2M100 model we could outperform the Google Translate, Microsoft Translator and eTranslation.

Keywords: natural language processing, multilingual model, sentiment analysis, abstractive summarization, machine translation, Marian NMT, M2M100

AMS Subject Classification: 68T07, 68T09, 68T50

Submitted: May 24, 2022
Accepted: November 2, 2022
Published online: November 9, 2022

1. Background

Several efforts have been made to analyze the tremendous amount of data that is currently available with the long-term goal to understand and analyze patterns. A highly promising approach towards that direction is the creation of generative models, that can generate new data instances similar to the original dataset. Recent advancements in artificial intelligence promote the development of systems with generative ability.

One aim of this research is to facilitate the work of administrators by processing human language. The members of the consortium that established the Infocommunication and Information Technology National Laboratory (ICT & IT National Laboratory) (the National Security Service and IdomSoft Zrt.) have set a dual goal: to support the safe introduction and use of emerging infocommunication and information technologies and the digital transformation of public administration.

One of IdomSoft LLC.'s[1] key objectives is to research and apply the potential of Artificial Intelligence (AI) based technologies for public administration applications, enabling customers to be exempted from the provision of all data already available in public administrations. The developments will save customers from all the organisational and administrative tasks that can be solved by internal administrative organisation between public administrations. The aim is also to create a secure and seamless contactless, fully digitised and automated administration.

This strategic innovation includes, among other things, the feasibility of public administration services that can handle the specificities of the Hungarian language at a high level of proficiency and meet the expectations of the 21st century. In order to achieve these objectives, IdomSoft LLC. cooperates with Hungarian universities to apply their products, which have been implemented in the R&D process, in practice in connection with the public administration IT solutions it develops.

Neural Machine Translation (NMT) is an important task in the area of Natural Language Processing (NLP), which is clearly highlighted by the fact that there is an increasing demand from the side of both academic and industrial stakeholders to push the limits of model performance and to come up with new, resource-efficient solutions. It is getting increasingly important to establish multilingual models that are able to handle dozens or even more than hundred languages simultaneously. The implementation of these multilingual models in certain directions and their application to NLP tasks in novel settings can promote the progress of machine translation in medium- or low-resourced languages.

Transfer learning represents a key strategy in enhancing model performance. It offers a solution to exploit the capabilities of a model that is trained for a certain task in order to use this knowledge to tackle other related problems. For example, cross-lingual knowledge transfer can substantially increase abstractive summmarization quality.

Our major research focus is to train multilingual models to NLP tasks followed by fine-tuning to specific tasks like text classification and abstractive summariza-

[1] https://idomsoft.hu

tion. We apply cross-lingual knowledge transfer to investigate how it can enhance model performance in our experimental settings.

Here we report that we could achieve highly superior performance with the models when cross-lingual knowledge transfer was applied. This further confirms that the application of transfer learning principles in NLP tasks can represent an outstanding opportunity to boost model performance and to establish competitive new approaches in the field of multilingual natural language processing.

2. Related work

The BART [16] is a transformer model developed by Fairseq (Facebook AI Research Sequence-to-Sequence Toolkit). The architecture of BART is based on two types of Transformers: the bidirectional encoder and the auto-regressive decoder. BART can be seen as a hybrid of a BERT- [8] and a GPT-type model [24]. The combination of the different features makes BART especially powerful and offers a unique opportunity to apply it for various purposes. For example, BERT models achieve impressive results in word- and sentence-level classification, while GPT models are well-suited for text generation tasks, such as summarization. BART can be applied with high success in machine translation, since it brings together the advantageous properties of both BERT-based and auto-regressive models.

The mBART (multilingual BART) is based on the seq2seq concept and it is a denoising autoencoder model pre-trained on corpora in multiple languages [20]. The application of mBART can significantly enhance the performance of both supervised and unsupervised machine translation, which can be especially promising in the case of translation of low- or medium-resourced languages. The mBART follows a sequence-to-sequence Transformer architecture [33] with 12 encoder and 12 decoder layers completed with an additional normalization layer. The conceptual framework of mBART is based on multilingual pre-training followed by fine-tuning to given language pairs. To pre-train the model, the CC25 corpus was applied [7] [15], which is a dataset consisting of 25 languages from different families. The texts were extracted from the CommonCrawl database and went through tokenization as a pre-processing step. The application of mBART could significantly improve the quality of both sentence-level and document-level machine translation, for example, in the case of low resource language pairs like English-Vietnamese or English-Turkish, more than 12 BLEU gains could be reached. On the contrary, for high resource language pairs, this performance gain was not observable or even resulted in a slightly worse performance. The results acquired by seq2seq-based approaches represent a significant improvement in the area of machine translation in comparison to previous efforts [20] [17]. The mBART was later expanded to mBART50 by incorporating additional 25 languages in the pipeline (doubling the number of the included languages), which resulted in remarkable BLEU improvements (up to 15 BLEU improvement in the case of some low resource languages) [30]. Taken together, the performance enhancement observed using mBART models suggests that there is transfer learning potential from the

representations acquired during multilingual pre-training. The mBART does not contains Hungarian language knowledge, thus we have pre-trained own English-Hungarian bilingual BART models.

Cross-lingual knowledge transfer can significantly improve model performance. For instance, Kahla et al. pre-trained a multilingual BERT model on a Hungarian corpus, then fine-tuned for abstractive summarization in Arabic. Similarly, the learned representations from pre-training on English corpus were transferred to Arabic in an attempt to improve the quality of summarization. The results indicate that it is possible to significantly elevate the quality of abstractive summarization by applying multilingual models pre-trained on a given language and transfer the acquired knowledge to another language [14]. The work by Artetxe et al. revealed important insights into the generalization ability of multilingual models and found that these models could achieve outstanding results on cross-lingual transfer benchmarks [2]. Additionally, cross-lingual knowledge transfer has been applied successfully in a variety of different areas, such as temporal expression extraction [6], name entity recognition [11], and utterance interpretation [26].

In machine translation research, there are only a few examples of multilingual models that can translate from any languages to Hungarian and vice versa. For research purposes, M2M100 [10] contains many languages including Hungarian, but it is an English-Centric model and it cannot translate from different languages to Hungarian. Among the industrial solutions, there are some multilingual translation systems, for instance Google Translate, Microsoft Translator or eTranslation, which use multilingual or bilingual models to translate from different languages to Hungarian.

The M2M100 project aimed at developing a translation tool comprising 103 different languages and 204 translation directions. A key proposition of the project was to initiate a paradigm shift in machine translation from English-Centric approaches towards multilingual model-based solutions [1]. Machine translation from multiple languages to multiple languages requires large datasets. This gave rise to a series of improvements in the generation of repositories with large data volume, including data mining [3] and reverse translation [28]. Hungarian translation capability is covered in M2M100, therefore it can be exploited in our projects as well.

The Marian NMT framework [13] is written in C++ language, which is an easy to install and well-annotated machine translation tool. Furthermore, its efficiency regarding memory usage and resource requirements makes it especially competent. Additionally, its minimal dependency on other technical solutions facilitates its application on a wider scale [12]. Due to its highly advantageous characteristics, Marian NMT is the most commonly used machine translation tool by academic users and developers [4]. Marian NMT operates using an attention model supported by an encoder-decoder architecture. Marian NMT is based on a neural machine translation model and it can reach the fastest runtime learning without the use of pre-training. In our experiment, a Marian large model was trained with the following specifications: 6 encoder layers and 6 decoder layers; 16 heads of at-

tention; words embedding dimension: 1024; input length: 1024 token; pre-attached mesh size: 4096.

The Google Translate [35] was launched in 2003. During the first phase, its operating principle was restricted to statistical machine translation, which was superseded by neural network-based machine translation in 2016. The quality of the translation has been significantly improved with the introduction of the neural network-based approach. This largely affected the performance in terms of inferences on a broader context and consequently more authentic translations. The Google Translate provides a record of results with several types of different translated versions, for example in the case of languages with gender distinction (e.g. French or Spanish), the feminine version is listed first followed by the masculine version [25]. Google Translate has the ability to handle 109 different languages with the add-on feature of translating spoken texts since 2020.

The Bing Translator is a machine translation solution developed by Microsoft Cognitive Services. It is capable of translating texts in more than 100 different languages and it even provides a solution for translating entire documents. Initially, it applied statistical machine translation, which was replaced by a neural network-based approach in 2018. Xu Tan et al. have developed a tool [29] to overcome the difference in the accuracy between multilingual and monolingual models, which is based on the knowledge distillation principle [5]. The core principle behind knowledge distillation is to increase efficiency and model performance by designating a 'student model', that can achieve the performance of a 'teacher model' or a set of models. The way this concept is implemented to machine translation means that there are language pair-specific teacher models that are used to train the student model that acquire the capability of handling all the languages by the teacher models. The effectiveness of this methodology is represented by its advanced performance in translation of TED talk transcripts from 44 languages to English, during which a BLEU-score improvement of 1 or even higher was achieved [29].

eTranslation[2] is an automated translation solution that can be applied to translate texts or entire documents written in any of the official languages of the Member States of the European Union, as well as Icelandic, Norwegian, Russian and simplified Chinese. The aim of the European Commission with the launch of eTranslation was to support small and medium-sized companies in the European Union, moreover to facilitate the interaction between public service providers, administrative officials and SMEs. The eTranslation tool can be especially useful, when translation capability is required during administrative and bureaucratic tasks. It is important to highlight that it can be easily integrated with other supporting digital solutions. To further support the machine translation procedure, several processing steps and text filtering options are also available under the CEF eTranslation Building Block project. A good example of that is the built-in option, which first divides long sentences into smaller parts before translation, which are later reconstructed to a coherent text. The eTranslation system has been trained on texts with subject-

[2] https://ec.europa.eu/info/resources-partners/machine-translation-public-administrations-etranslation_en

specific content, such as tenders, legal and medical texts, etc. The model has been trained in 24 different languages on more than 1 billion sentences.

3. Corpora

In order to train our bilingual BART models, two different corpora were used: Hungarian and English Wikipedia. In Table 1, you can see the characteristics of the two corpora.

Table 1. Characteristics of the pre-training corpora for BART.

	Segment	Token	Type	Paragraph sentence # (median)	Paragraph token # (median)
English WikiText-103	707,391	96,534,563	596,820	5	125
Hungarian Wikipedia	1,098,156	90,349.849	3,137,980	4	69

For fine-tuning our BART models to sentiment analysis task, we used the Hungarian Twitter Sentiment Corpus[3] that is created by Precognox[4]. According to the international benchmarks [34] we created two subcorpora from this corpus:

- 2-class (HTS2): binary classification subcorpus. We have converted the scores 1 and 2 to 0 as negative, scores 4 and 5 to 1 as positive. Score 3 was ignored to avoid the ambiguities. Training corpus: 2,468 segments. Test corpus: 269 segments.
- 5-class (HTS5): original five-point likert scaled corpus. 1: very negative, 2: negative, 3: neutral, 4: positive, 5: very positive. Training corpus: 3,600 segments. Test corpus: 400 segments.

For the zeroshot and transfer sentiment analysis experiments, we used the SST2 and SST5 corpora from GLUE [34] benchmark.

For the summarization task, we used the H+I corpus that Yang et al. used in their research [36], NOL (Népszabadság online corpus; nol.hu online articles (art) and its' leads from 1999 to 2016) and MARCELL [32] (law documents (doc) and its' one line descriptions (desc) from 1991 to 2019) corpora. Table 2 shows the characteristics of the fine-tuning corpora. For the zeroshot and transfer summarization experiments, we used the CNN/Daily Mail [27] corpora.

In our machine translation task, we built Hungarian-Centric translation models with 12 languages, which means the source text can be in 12 different languages and the target language is Hungarian (hu) in all cases. The 12 different source languages are the following:

[3] http://opendata.hu/dataset/hungarian-twitter-sentiment-corpus
[4] https://www.precognox.com

Table 2. Characteristics of the fine-tuning corpora.

	Segment	Token #	Type #	Avg. token #
HTS2	2,737	42,797	13,713	15.62
HTS5	4,000	59,997	18,423	14.99
H+I	559,162	147,099,485 (art) 16,699,600 (lead)	2,949,173 (art) 749,586 (lead)	263.07 (art) 29.87 (lead)
NOL	397,343	153,003,164 (art) 15,786,166 (lead)	2,482,398 (art) 623,445 (lead)	384.52 (art) 39.71 (lead)
MARCELL	24,747	27,834,358 (doc) 277,732 (desc)	444,352 (doc) 29,189 (desc)	1124.82 (doc) 11.59 (desc)

- Bulgarian (bg), Czech (cs), German (de), English (en), Croatian (hr), Polish (pl), Romanian (ro), Russian (ru), Slovak (sk), Slovene (sl), Serbian (sr), Ukrainian (uk)

In order to build a machine translation from scratch, a huge amount of data is required. In contrast, for fine-tuning task, smaller amount of data is enough. Thus, we created two corpora for our task. First one contains 8 million (8M) segments per language (except for Ukrainian, due to lack of data it contains only 5,805,144 segments), the second one is a sub-corpus of the 8M corpus that contains 3 million (3M) segments for each language. The data were collected from OPUS [31] that is composed of the following sub-corpora:

- Bible, Bible-uedin, Books, CCAligned, CCMatrix, DGT, ECB, ELITR, ELITR-ECA, ELRC_2922, ELRC_2923, ELRC_3382, EMEA, EUbookshop, EUconst, Europarl, GNOME, GlobalVoices, JRC, JRC-Acquis, KDE4, KDEdoc, MultiCCAligned, MultiParaCrawl, OpenSubtitles, PHP, ParaCrawl, QED, TED2020, Tatoeba, TildeMODEL, Ubuntu, WMT-News, WikiMatrix, Wikimedia, Wikipedia, XLEnt

The different language pairs contain different composite of the sub-corpora. In Table 3, you can see the characteristics of the training sub-corpora for the machine translation task.

4. Experiments

In our pre-training experiment, we have trained two bilingual BART models of different size:

- **BART-base**: base size BART model trained on English and Hungarian Wikipedia. Main hyper-parameters: 6 encoder layers and 6 decoder layers; 12 attention heads; word embedding dimensions: 768; input length: 512; 140 million parameters.

Table 3. Characteristics of the machine translation corpora.

	Token 8M / 3M	Type 8M / 3M	Avg. token / sent 8M / 3M
bg	101,701,016 / 38,149,260	998,060 / 586,926	12.71 / 12.72
hu	93,370,875 / 35,023,413	1,843,452 / 1,057,434	11.67 / 11.68
cs	96,854,637 / 36,345,169	1,369,081 / 797,557	12.11 / 12.12
hu	96,313,811 / 36,125,748	2,008,769 / 1,141,009	12.04 / 12.04
de	123,826,131 / 46,407,141	1,708,615 /957,634	15.48 / 15.47
hu	113,026,306 / 42,365,265	2,215,093 / 1,267,205	14.13 / 14.12
en	118,593,896 / 44,440,629	1,112,914 / 593,035	14.82 / 14.81
hu	104,287,145 / 39,072,921	2,375,910 / 1,331,924	13.04 / 13.02
hr	78,932,860 / 29,601,947	1,075,070 / 631,246	9.87 / 9.87
hu	78,540,254 / 29,445,821	1,685,025 / 961,367	9.82 / 9.82
pl	97,533,671 / 36,584,480	1,350,775 / 793,299	12.19 / 12.20
hu	98,984,434 / 37,126,013	2,062,157 / 1,166,764	12.37 / 12.38
ro	110,276,300 / 41,357,056	952,906 / 555,642	13.79 / 13.79
hu	93,431,714 / 35,058,265	1,906,878 / 1,091,748	11.68 / 11.69
ru	88,227,629 / 33,085,548	1,376,699 / 807,518	11.03 / 11.03
hu	85,205,960 / 31,956,481	1,838,741 / 1,049,578	10.65 / 10.65
sk	122,935,150 / 46,085,577	1,567,148 / 920,586	15.37 / 15.36
hu	123,016,834 / 46,105,105	2,225,916 / 1,278,686	15.38 / 15.37
sl	106,838,393 / 40,042,349	1,195,476 / 703,052	13.36 / 13.35
hu	106,714,770 / 40,013,573	1,973,244 / 1,138,862	13.34 / 13.34
sr	72,647,210 / 27,237,077	1,185,523 / 710,495	9.08 / 9.08
hu	71,058,803 / 26,642,218	1,446,568 / 832,887	8.88 / 8.88
uk	70,816,656 / 36,581,363	1,306,774 / 927,544	12.20 / 12.19
hu	69,564,268 / 35,933,267	1,556,554 / 1,088,340	11.98 / 11.98

- **BART-large**: large size BART model that trained on English and Hungarian Wikipedia. Main hyper-parameters: 12 encoder layers and 12 decoder layers; 16 attention heads; word embedding dimensions: 1024; input length: 1024; 400 million parameters.

In our fine-tuning experiments, we performed three different tasks:

1. **Sequence classification**: Using our pre-trained bilingual BART models and two multilingual BERT-based models (mBERT [9] and XLM-RoBERTa [19]), we carried out three different experiments in sentence-level sentiment analysis:

    - *baseline*: We fine-tuned and tested the four models on HTS2 and HTS5.
    - *zeroshot*: We fine-tuned the four models on SST2 and SST5, then tested on HTS2 and HTS5.

- *transfer*: We fine-tuned the four models on SST2 and SST5, then further fine-tuned on HTS2 and HTS5, finally tested on HTS2 and HTS5.

2. **Text summarization**: We fine-tuned the BART base model on three different corpora: H+I, NOL and MARCELL. Because of hardware limits, we could not fine-tuned our BART large model on summarization task. We carried out two different experiments in text summarization task:
   - *baseline*: We fine-tuned and tested our model on the three corpora.
   - *transfer (tf)*: We fine-tuned our model on CNN/Daily Mail, then further fine-tuned and tested on the three Hungarian corpora.
3. **Machine translation**: We fine-tuned the M2M100 large model (facebook/m2m100_1.2B[5]) on the 3M sub-corpus for machine translation. The source text can be in 12 different languages, the target text is Hungarian. In this experiment, we fine-tuned our model with only 1 epoch.
   - *From scratch*: In the case of machine translation, we also trained a multilingual translation model from scratch. For this task, we used the Marian NMT [13] framework. For training Marian NMT model, we used the 8M corpus for machine translation. Similar to the M2M100 experiment, the source text can be in 12 different languages, the target text is Hungarian. To help the translation model, we inserted the language code at the beginning of the source segments in the following format (lang is the ISO language code): ___lang___. A Marian large model was trained with 66 epoch.

5. Results

In order to evaluate our experiments, the following metrics were used:

- Accuracy: In the case of sentiment analysis tasks, accuracy metrics were used.
- ROUGE [18] : For summarization tasks, we used the ROUGE metrics in the following format: ROUGE-1/ROUGE-2/ROUGE-L.
- BLEU [21], chrF [22]: For word-level and character-level evaluation of machine translation, SacreBLEU [23] and chrF-6 metrics were used in the following format: BLEU/chrF-6.

In Table 4, you can see the results of the sentiment analysis experiments. For transfer and zeroshot tasks, first, we fine-tuned the models on the English SST corpora. Under the double line, you can see the results of the SST fine-tuning. Above the double line, you can see the results of our experiments. In all cases, the transfer task could increase the result of the models. It can prove that adding relevant data to model could increase performance, even if it is in another language.

[5] https://huggingface.co/facebook/m2m100_1.2B

Table 4. Sentiment analysis results.

	HTS2	HTS5
BART-base (baseline)	74.44	56.75
BART-base (zeroshot)	42.96	28.75
BART-base (transfer)	**74.81**	**57.25**
BART-large (baseline)	74.07	56.00
BART-large (zeroshot)	44.81	23.50
BART-large (transfer)	**74.59**	**56.74**
mBERT (baseline)	78.51	57.74
mBERT (zeroshot)	47.41	30.50
mBERT (transfer)	**80.37**	**57.99**
XLM (baseline)	83.33	63.49
XLM (zeroshot)	68.88	40.99
XLM (transfer)	**84.81**	**79.79**
	SST2	SST5
BART-base	79.01	36.72
BART-large	80.27	36.36
mBERT	90.59	49.57
XLM	93.34	50.43

In Table 5, you can see the results of the summarization task. Similar to the classification task, under the double line, you can see the result of the fine-tuning on the English CNN/Daily Mail corpora. Above the double line, you can see our experiment. As you can see in the Table 5, transfer method in this case could also increase the performance.

Table 5. Abstractive summarization results.

	H+I	NOL	MARCELL
BART-base (baseline)	31.4/14.3/23.5	42.7/27.6/35.4	71.5/63.0/69.9
BART-base-tf	**31.8/14.5/23.5**	**45.1/30.5/37.6**	**77.1/70.6/76.0**
	CNN/Daily Mail		
BART-base	40.1/17.6/27.4		
BART en original	44.2/21.3/40.9		

In Table 6, you can see the results of the machine translation experiments. We have compared our Marian and M2M100 models with Google Translate, Microsoft Translator and eTranslation (the eTranslation cannot translate serbian, thus this results is missing).

The M2M100 fine-tuning results in significantly higher scores then any other tools included in our experimental analysis. Compared to Marian, M2M100 uses only 3 million segments for each language, and only 1 epoch for fine-tuning. It means, the model could transfer significant amount of knowledge from the pre-

Table 6. Comparision of performance of different machine translation models.

	Marian	**M2M100**	Google	Microsoft	eTranslation
bg	21.3/43.9	**26.6/48.0**	20.0/43.6	20.8/44.2	22.3/45.6
cs	22.5/46.0	**28.9/50.3**	22.6/45.3	23.1/45.9	24.7/47.4
de	21.9/46.2	**28.3/51.4**	22.7/48.0	22.8/47.8	24.0/48.8
en	27.7/49.6	**34.4/54.7**	25.3/49.1	26.3/50.3	28.3/51.3
hr	19.2/42.7	**26.2/47.3**	19.6/42.5	20.1/43.1	20.9/43.7
pl	21.2/45.2	**28.3/50.2**	21.4/45.4	22.2/45.7	23.9/47.2
ro	19.5/43.8	**26.4/48.7**	21.0/44.9	21.8/45.7	23.6/46.9
ru	19.7/43.9	**25.1/48.1**	19.8/44.5	21.0/45.6	20.3/44.8
sk	23.1/48.9	**30.9/53.9**	22.6/47.7	23.1/48.5	26.4/50.8
sl	22.7/45.8	**27.7/50.5**	14.4/34.4	21.5/45.0	26.0/48.4
sr	18.0/40.5	**23.4/44.7**	18.0/40.7	19.2/41.5	-
uk	24.2/49.8	**32.6/55.2**	21.8/47.0	23.3/48.2	22.9/48.5
avg	21.8/45.5	**28.2/50.3**	20.8/44.4	22.1/46.0	23.9/47.6

trained 100 language. Thus, less amount data and training steps are enough to achieve higher results. Our Marian experiment used 2.5x larger corpora and 66x more epoch and still gained lower performance than our fine-tuned M2M100 model, but still better than the Google Translate, for instance. Our Marian model could not outperformed the eTranslation, which is not surprising, because the eTranslation uses different bilingual models to translate, and a bilingual model is more accurate than a 12-lingual model. Therefore, our M2M100 model is an outstanding result, because it uses only one model that can gain better results than the bilingual models.

6. Conclusion

In our research, we pre-trained and fine-tuned different transformer-based multilingual generative models for Hungarian natural language processing tasks. We have carried out four different experiments. For pre-training language model, encoder-decoder autoregressive BART models were applied. As classification task, we fine-tuned four different models for sentiment analysis. For summarization task, our pre-trained BART base model was fine-tuned on three different corpora. We also did experiments in zero-shot and cross-lingual transfer learning settings. Last but not least, we built the first (two at once) 12-lingual Hungarian-Centric machine translation model, which uses only one model to translate from 12 languages to Hungarian. In this task, we trained a model from scratch and the M2M100 model was fine-tuned. Our fine-tuned M2M100 used much less data and training steps and yet, it could outperform the Google Translate, the Microsoft Translator and the eTranslation.

Acknowledgements. The research reported in the current publication was carried out by affiliated members of the Pázmány Péter Catholic University and the IdomSoft Ltd, and it was supported by the Ministry of Innovation and Technology and the National Research, Development and Innovation Office within the framework of the National Laboratory of Infocommunication and Information Technology.

References

[1] R. Aharoni, M. Johnson, O. Firat: *Massively Multilingual Neural Machine Translation*, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 3874–3884, doi: https://doi.org/10.18653/v1/N19-1388.

[2] M. Artetxe, S. Ruder, D. Yogatama: *On the Cross-lingual Transferability of Monolingual Representations*, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, doi: https://doi.org/10.18653/v1/2020.acl-main.421.

[3] M. Artetxe, H. Schwenk: *Massively Multilingual Sentence Embeddings for Zero-Shot Cross-Lingual Transfer and Beyond*, Transactions of the Association for Computational Linguistics 7 (2019), pp. 597–610, doi: https://doi.org/10.1162/tacl_a_00288.

[4] L. Barrault, O. Bojar, M. R. Costa-jussà, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, P. Koehn, S. Malmasi, C. Monz, M. Müźller, S. Pal, M. Post, M. Zampieri: *Findings of the 2019 Conference on Machine Translation (WMT19)*, in: Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1), Florence, Italy: Association for Computational Linguistics, 2019, pp. 1–61.

[5] C. Bucila, R. Caruana, A. Niculescu-Mizil: *Model Compression*, in: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '06, Philadelphia, PA, USA: Association for Computing Machinery, 2006, pp. 535–541, isbn: 1595933395.

[6] Y. Cao, W. Groves, T. K. Saha, J. R. Tetreault, A. Jaimes, H. Peng, P. S. Yu: *XLTime: A Cross-Lingual Knowledge Transfer Framework for Temporal Expression Extraction*, in: arXiv, 2022, doi: https://doi.org/10.48550/ARXIV.2205.01757.

[7] P.-J. Chen, J. Shen, M. Le, V. Chaudhary, A. El-Kishky, G. Wenzek, M. Ott, M. Ranzato: *Facebook AI's WAT19 Myanmar-English Translation Task Submission*, 2019, doi: https://doi.org/10.48550/ARXIV.1910.06848.

[8] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova: *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186, doi: https://doi.org/10.18653/v1/N19-1423.

[9] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova: *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186.

[10] A. Fan, S. Bhosale, H. Schwenk, Z. Ma, A. El-Kishky, S. Goyal, M. Baines, O. Çelebi, G. Wenzek, V. Chaudhary, N. Goyal, T. Birch, V. Liptchinsky, S. Edunov, E. Grave, M. Auli, A. Joulin: *Beyond English-Centric Multilingual Machine Translation*, ArXiv abs/2010.11125 (2020).

[11] X. Feng, X. Feng, B. Qin, Z. Feng, T. Liu: *Improving Low Resource Named Entity Recognition using Cross-lingual Knowledge Transfer*, in: July 2018, pp. 4071–4077, doi: https://doi.org/10.24963/ijcai.2018/566.

[12] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. F. Aji, N. Bogoychev, A. F. T. Martins, A. Birch: *Marian: Fast Neural Machine Translation in C++*, in: Proceedings of ACL 2018, System Demonstrations, Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 116–121, doi: https://doi.org/10.18653/v1/P18-4020, url: https://aclanthology.org/P18-4020.

[13] M. Junczys-Dowmunt, R. Grundkiewicz, T. Dwojak, H. Hoang, K. Heafield, T. Neckermann, F. Seide, U. Germann, A. Fikri Aji, N. Bogoychev, A. F. T. Martins, A. Birch: *Marian: Fast Neural Machine Translation in C++*, in: Proceedings of ACL 2018, System Demonstrations, Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 116–121.

[14] M. Kahla, Z. G. Yang, A. Novák: *Cross-lingual Fine-tuning for Abstractive Arabic Text Summarization*, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), Held Online: INCOMA Ltd., Sept. 2021, pp. 655–663.

[15] G. Lample, A. Conneau: *Cross-lingual Language Model Pretraining*, 2019, doi: https://doi.org/10.48550/ARXIV.1901.07291.

[16] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer: *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online: Association for Computational Linguistics, July 2020, pp. 7871–7880, doi: https://doi.org/10.18653/v1/2020.acl-main.703.

[17] X. Li, G. Li, L. Liu, M. Meng, S. Shi: *On the Word Alignment from Neural Machine Translation*, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy: Association for Computational Linguistics, July 2019, pp. 1293–1303, doi: https://doi.org/10.18653/v1/P19-1124.

[18] C.-Y. Lin: *ROUGE: A Package for Automatic Evaluation of Summaries*, in: Text Summarization Branches Out, Barcelona, Spain: Association for Computational Linguistics, July 2004, pp. 74–81.

[19] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov: *RoBERTa: A Robustly Optimized BERT Pretraining Approach*, CoRR (2019).

[20] L. Miculicich, D. Ram, N. Pappas, J. Henderson: *Document-Level Neural Machine Translation with Hierarchical Attention Networks*, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium: Association for Computational Linguistics, 2018, pp. 2947–2954, doi: https://doi.org/10.18653/v1/D18-1325.

[21] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu: *Bleu: a Method for Automatic Evaluation of Machine Translation*, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, July 2002, pp. 311–318, doi: https://doi.org/10.3115/1073083.1073135.

[22] M. Popović: *chrF: character n-gram F-score for automatic MT evaluation*, in: Proceedings of the Tenth Workshop on Statistical Machine Translation, Lisbon, Portugal: Association for Computational Linguistics, Sept. 2015, pp. 392–395, doi: https://doi.org/10.18653/v1/W15-3049.

[23] M. Post: *A Call for Clarity in Reporting BLEU Scores*, in: Proceedings of the Third Conference on Machine Translation: Research Papers, Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 186–191, doi: https://doi.org/10.18653/v1/W18-6319.

[24] A. Radford, K. Narasimhan: *Improving Language Understanding by Generative Pre-Training*, in: 2018.

[25] A. A. Rescigno, J. Monti, A. Way, E. Vanmassenhove: *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*, in: Workshop on the Impact of Machine Translation (iMpacT 2020), Virtual: Association for Machine Translation in the Americas, Oct. 2020, pp. 62–90.

[26] S. Schuster, S. Gupta, R. Shah, M. Lewis: *Cross-lingual Transfer Learning for Multilingual Task Oriented Dialog*, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 3795–3805, doi: https://doi.org/10.18653/v1/N19-1380.

[27] A. See, P. J. Liu, C. D. Manning: *Get To The Point: Summarization with Pointer-Generator Networks*, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, Canada: Association for Computational Linguistics, July 2017, pp. 1073–1083, doi: https://doi.org/10.18653/v1/P17-1099.

[28] R. Sennrich, B. Haddow, A. Birch: *Neural Machine Translation of Rare Words with Subword Units*, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1715–1725, doi: https://doi.org/10.18653/v1/P16-1162, url: https://aclanthology.org/P16-1162.

[29] X. Tan, Y. Ren, D. He, T. Qin, T.-Y. Liu: *Multilingual Neural Machine Translation with Knowledge Distillation*, in: International Conference on Learning Representations, 2019.

[30] Y. Tang, C. Tran, X. Li, P.-J. Chen, N. Goyal, V. Chaudhary, J. Gu, A. Fan: *Multilingual Translation with Extensible Multilingual Pretraining and Finetuning*, 2020, doi: https://doi.org/10.48550/ARXIV.2008.00401.

[31] J. Tiedemann: *Parallel Data, Tools and Interfaces in OPUS*, in: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), ed. by N. C. (Chair), K. Choukri, T. Declerck, M. U. Dogan, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, Istanbul, Turkey: European Language Resources Association (ELRA), 2012, isbn: 978-2-9517408-7-7.

[32] T. Váradi, S. Koeva, M. Yamalov, M. Tadić, B. Sass, B. Nitoń, M. Ogrodniczuk, P. Pęzik, V. Barbu Mititelu, R. Ion, E. Irimia, M. Mitrofan, V. Păiș, D. Tufiș, R. Garabík, S. Krek, A. Repar, M. Rihtar, J. Brank: *The MARCELL Legislative Corpus*, English, in: Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France: European Language Resources Association, May 2020, pp. 3761–3768, isbn: 979-10-95546-34-4.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin: *Attention is All you Need*, in: Advances in Neural Information Processing Systems 30, ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett, Curran Associates, Inc., 2017, pp. 5998–6008.

[34] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S. Bowman: *GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding*, in: Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 353–355, doi: https://doi.org/10.18653/v1/W18-5446.

[35] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, J. Dean: *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*, CoRR abs/1609.08144 (2016).

[36] Z. G. Yang, Á. Agócs, G. Kusper, T. Váradi: *Abstractive text summarization for Hungarian*, Annales Mathematicae et Informaticae 53 (2021), pp. 299–316.