Evaluation of Transfer Learning for Polish with a text-to-text model

Anonymous ACL submission

Abstract

We present polT - a general purpose text-to-text model for Polish that can be fine-tuned on a variety on Natural Language Processing (NLP) tasks with a single training objective. Unsupervised denoising pre-training is performed efficiently by initializing the model weights with multi-lingual T5 (mT5) counterpart. We evaluate performance of polT, mT5, Polish BART (plBART) and Polish GPT-2 (papuGaPT2) on diverse downstream tasks such as: text-to-text KLEJ benchmark, en-pl machine translation, question answering and summarization. The polT scores top on all of this tasks except summarization where plBART is best. In general (except summarization), the larger the model the better the results. The encoder-decoder architectures prove to be better than decoder-only equivalent. Additionally, since summarization and question answering lack benchmark datasets for Polish language we describe in detail their construction and will make them publicly available.

1 Introduction

Recent years have brought significant progress in both natural language understanding (NLU) and natural language generation (NLG). Transformer architecture enabled efficient training of large-scale language models (Radford et al., 2019) and language understanding models (Devlin et al., 2019; Liu et al., 2019). On the other hand, transfer learning, which has been used in representation learning for years (Mikolov et al., 2013; Pennington et al., 2014; Devlin et al., 2019), has finally been successfully applied to text-to-text problem as well (Raffel et al., 2020; Lewis et al., 2020). The text-to-text framework takes text as input and produces new text as output. This unified view enables the use of same architecture, training procedure and decoding process for many NLP tasks such as classification, machine translation, summarization and question answering, to name a few. Raffel et al. (2020) demonstrated that simplicity of this approach combined with scale can achieve state-of-the-art results on many benchmark datasets. Their T5 model was available only for English language, but more recently pre-trained multi-lingual architectures (Liu et al., 2020; Xue et al., 2021; Nagoudi et al., 2021a) and non-English counterparts (Carmo et al., 2020; Husein, 2018) were released. It is now known that models targeted for specific language perform better than multi-lingual one (Martin et al., 2020; Le et al., 2020; Chan et al., 2020; Mroczkowski et al., 2021; Virtanen et al., 2019; Nagoudi et al., 2021b). Moreover, specialized architectures are typically smaller due to significantly reduced vocabulary size and they can be trained efficiently via transfer from multi-lingual checkpoints (Arkhipov et al., 2019; Mroczkowski et al., 2021). There were some attempts to pre-train Transformer-based models for generating Polish namely plBART [1] and papuGaPT2 [2], but they lack detailed description and evaluation on benchmark datasets.

Our contributions are:

1. demonstration of the efficiency of pre-training procedure for transferring knowledge from multi-lingual to monolingual text-to-text models based on work by Arkhipov et al. (2019); Mroczkowski et al. (2021),

2. comprehensive evaluation of text-to-text models on diverse tasks in Polish, such as text-to-text KLEJ benchmark (Rybak et al., 2020), machine translation, question answering and summarization

3. release of polT[3] – a T5-based model for the Polish language, which achieves the best[4] re-

[1] https://github.com/sdadas/polish-nlp-resources#bart

[2] https://huggingface.co/flax-community/papuGaPT2

[3] https://proper.download.site/model

[4] among text-to-text models

sults on KLEJ benchmark, machine translation and question answering and second best results in summarization.

4. construction of benchmark datasets in Polish domain for question answering and summarization.

2 Polish T5 model

There was no large-scale text-to-text model for Polish to benchmark transfer learning methods, so we trained 3 versions of polT model (small, base, and large) with T5 architecture. We prepared a Polish language tokenizer, initialized the model weights based on mT5, and pre-trained it on Polish corpora.

2.1 Architecture

We followed the original mT5 (Xue et al., 2021) architectures with encoder and decoder stacks for all trained model sizes. We trained 3 variants of the T5 model: small (8 layers, 6 attention heads, hidden dimension 1024), base (12 layers, 12 attention heads, hidden dimension 768) and large (24 layers, 16 attention heads, hidden dimension 1024).

2.2 Initialization

Parameters for all layers except word embeddings were copied directly from public mT5 models (Xue et al., 2021). The difference in tokenizers vocabulary was addressed as described in (Arkhipov et al., 2019; Mroczkowski et al., 2021). Embedding weights for tokens appearing in both dictionaries were copied directly. Otherwise, if target vocabulary token was missing in source vocabulary it was split into sub-tokens and the embedding was the average of sub-tokens embeddings.

2.3 Pre-training datasets

Denoising pre-training was performed on a weighted mixture[5] of the following corpora: Polish Wikipedia, National Corpus of Polish (NKJP, Przepiórkowski et al. 2011), Wolne Lektury[6], Polish Open Subtitles[7] and Common Crawl of Polish sites. Each dataset is described in Mroczkowski et al. (2021).

[5]Probability of selecting given corpus in the mixture was proportional to the size of the dataset.

[6]https://wolnelektury.pl/

[7]https://www.opensubtitles.org/pl

2.4 Tokenizer

The training dataset was tokenized into subword units using a sentencepiece unigram model (Kudo, 2018) with vocabulary size of 50k tokens. Unigram model was trained[8] on the most representative parts of our corpus, i.e subset (quarter of the whole corpus) of the NKJP, and Polish Wikipedia.

2.5 Pre-training procedure

For unsupervised denoising pre-training we used the original T5 training scripts[9], which are based on the Mesh TensorFlow framework (Shazeer et al., 2018). We used default options for span-corruption objective with a mean span length of 3 and corruption rate of 15% (Raffel et al., 2020).

polT models were trained using AdaFactor optimizer (Shazeer and Stern, 2018) with constant warmup and inverse square root learning rate schedule (Raffel et al., 2020, section 3.1.2). We use the peak learning rate 5e-3, batch of 1048576 tokens, 0% dropout, and trained for 50k steps with 1024 warmup steps on a single preemptible TPU v3.

The learning rate was found in simple hypertuning procedure. We evaluated polT-base model snapshots saved after 10k training steps for pre-training learning rates in {1e-3, 2e-3, 5e-3}. For performance evaluation we used KLEJ benchmark tasks equipped with validation sets, namely: AR, NKJP-NER, PolEmo2.0-OUT, CDSC-E. Best model training was continued up to 50k steps.

3 Downstream tasks

polT is a generic text-to-text model for Polish. It can be fine-tuned on many NLP tasks. We evaluated it by fine-tuning on text-to-text reformulation of KLEJ benchmark, machine translation, question answering and summarization. The same fine-tuning scheme was used for mT5 and other text-to-text or NLG models for Polish: plBART[10] and papuGaPT2[11], in order to compare their performance. Trainable number of parameters and vocabulary size for each architecture is show in Table 1.

[8]https://github.com/google/sentencepiece

[9]https://github.com/google-research/text-to-text-transfer-transformer

[10]https://github.com/sdadas/Polish-nlp-resources#bart

[11]https://huggingface.co/flax-community/papuGaPT2

Model		# Params	Vocab. size
Small models	mT5	300M	250k
	polT	95M	50k
Base models	mT5	582M	250k
	polT	275M	50k
	plBART	139M	50k
	papuGaPT2	124M	50k
Large models	mT5	1.23B	250k
	polT	820M	50k

Table 1: Summary of total number of trainable parameters for pre-trained models for Polish.

3.1 General language understanding

We verified the quality of the assessed models in terms of general language understanding with KLEJ benchmark (Rybak et al., 2020). To compare the generative models, we used all of the given tasks, except for the CDSC-R regression. We formed the tasks into a text-to-text format (see Appendix A) proposed in (Raffel et al., 2020). During fine-tuning phase, the model was provided descriptive input containing specific prefix and the textual features with descriptive labels[12]. The training objective was to generate greedily a specific label consisting of up to several tokens. Targets were generated over the whole vocabulary and only an exact match was treated as the correct answer.

3.1.1 Experiments

To ensure a fair comparison between models, we followed a simple experimental setup. The learning rate was the only hyperparameter we tuned (see Appendix A.3). By default, we trained the models for 1600 steps with a batch size of 64. Reported results are the median of 7 runs for small and base architectures and the median of 3 runs for the large one.

3.1.2 Results

The results for KLEJ benchmark (text-to-text format) are in Table 2. In general, encoder-decoder models dedicated to the Polish language perform significantly better at tasks in the KLEJ benchmark than the multi-lingual counterparts. The most striking performance difference occurs for the base version of the model on the PolEmo2.0-OUT task. The difference between polT-base and mT5-base is over 9 pp. The difference on PolEmo2.0-OUT of more than 2pp is also visible for the other sizes. We hypothesize that domain adaptation necessary to solve this task reveals better Polish language understanding for monolingual T5. The superiority of the dedicated model is also visible in the PSC paraphrase identification task. Especially for the T5-small and T5-base models, the difference is 4.6pp and 3.9pp, respectively. The gap closes to 1pp for the large version of a T5 model. On the other hand, the multi-lingual mT5-large model obtained the best result on the NKJP-NER task, which shows that the task formulated in a classification form is more straightforward than its original version.

We observe a consistent increase in results with the size of the polT model in each task. The best encoder-decoder model is polT-large. It performs the best in 6 out of the 8 evaluated tasks from the KLEJ benchmark. However, it is not as good as the best Polish encoder, namely HerBERT-large. The difference between these architectures is, on average, 1pp. We observe the most significant performance gap of 6.8pp when moving from polT-small to polT-base. We found the biggest difference of around 20pp on *PSC* and *Czy wiesz?* tasks.

Regarding the smaller architectures, the best text-to-text model is polT-base. It is worth emphasizing that almost 2 times smaller plBART is very competitive. It is the best on 3 out of 8 tasks among the smaller models. The decoder only papuGaPT2 achieves on average results on pair with polT-small but is the worst at dealing with the highly unbalanced tasks, *CBD* and *Czy wiesz?*.

We emphasize the variability of the results for the two *CBD* and *Czy wiesz?* tasks. The difference between the results for different architectures is as high as 30pp.

To sum up, the models dedicated to the Polish language are the best from the perspective of the KLEJ benchmark. Among them, the T5 models were the best.

3.2 Summarization

Summarization is a task of writing a shorter text that has all the key information of the longer text. Summaries can vary in length and be extractive (a selection of passages from the original text) or abstractive (written from scratch, though passages from the orignal text are not forbidden).

[12]We didn't explore multi-task learning in this setup with task-specific prompts

Model		AVG	NKJP-NER	CBD	Czy wiesz?	PolEmo2.0-IN	PolEmo2.0-OUT	AR	PSC	CDSC-E
Small models	mT5	74.3 ± 2.6	88.7	56.6	42.7	86.0	73.2	84.6	70.8	91.9
	polT	76.8 ± 1.8	92.4	60.4	44.7	88.0	76.6	84.7	75.4	92.5
Base models	mT5	82.4 ± 0.0	92.8	65.6	67.8	88.4	70.2	87.6	93.3	93.1
	papuGaPT2	76.5 ± 0.9	90.7	33.3	49.8	89.2	76.2	86.2	95.3	91.3
	plBART	81.9 ± 0.5	93.1	47.6	68.5	89.5	77.9	88.0	97.9	93.2
	polT	83.4 ± 0.3	93.6	62.3	63.8	90.0	79.3	87.8	97.2	93.4
	HerBERT	84.7 ± 0.4	94.5	66.4	64.3	90.9	80.4	87.7	98.9	94.5
Large models	mT5	84.9 ± 6.7	**94.8**	62.3	**69.9**	91.4	80.3	88.8	97.9	93.5
	polT	**86.4 ± 0.3**	94.5	**70.0**	69.4	**92.9**	**82.6**	**88.9**	**98.9**	**94.0**
	HerBERT	87.5 ± 0.2	96.4	72.0	75.8	92.2	81.8	89.1	98.9	94.1

Table 2: Evaluation results on KLEJ benchmark. *AVG* is the average score across all tasks. Scores are reported for test set and correspond to median values across 7 runs for small and base models and 3 runs for large version of models. The best scores for text-to-text models within each group are underlined, the best overall are in bold. Scores for the HerBERT model were taken from (Mroczkowski et al., 2021) and evaluated in a standard setting.

3.2.1 Datasets

Allegro Articles (AA) Collection of over 33k articles from Allegro.pl[13] - a popular e-commerce marketplace. They contain among others product reviews and shopping guides. Every article contains a title, lead (opening paragraph) and a body of text. We prepared 2 different summarization tasks: (1) *body2lead* - generate lead from a body of the article (2) *body+lead2title* - generate a title from a full article (lead and body). The tasks are entirely abstractive and differ in source and target length. The details of dataset construction and statistics are in Appendix B.1.

Polish Summaries Corpus (PSC) Collection of summaries for 569 news articles created by human annotators (Ogrodniczuk and Kopeć, 2014). The annotators created 5 extractive summaries (each by different author) for each article by choosing approximately 5%, 10%, or 20% of the original text. The subset of 154 articles was also supplemented with additional 5 abstractive summaries each, i.e., not created from the fragments of the original article. We constructed 3 summarization subsets: (1) *whole* - all summaries and articles (2) *extract* - only extractive summaries and (3) *abstract* - only abstractive summaries. The details of dataset construction and statistics are in Appendix B.2.

[13] https://allegro.pl/artykuly

3.2.2 Experiments

For a fair comparison between models, we designed a common experimental setup, where the learning rate is the only hyperparameter (details in Appendix B.4). By default, each model was trained for 10 epochs with a batch size of 32 and Adam optimizer (Kingma and Ba, 2014). The only exceptions were large versions of T5 models (all tasks) and papuGaPT2 (only on PSC task), which were trained for 4 epochs. Each experiment, including hyperparameter search, was repeated with 3 seed.

Trimming inputs Summarization models by its nature deal with long source texts and shorter target texts. Unfortunately, we had to substantially trim source texts due to model and memory limitations. Usually, the limit was 1024 tokens, but sometimes it was even lower. The PSC tasks (news articles domain) were affected the most. None of the models ever saw 100% of source text during training, the percentage varied between 26-52% depending on the model. Details can be found in Appendix B.3.

3.2.3 Results

Results are shown in Table 3 which contains arithmetic mean of (f-measure) ROUGE-1, ROUGE-2 and ROUGE-L (Lin, 2004)[14] for each model and task. More detailed metrics for each tasks can be found in Appendix D (Tables 10-14). In all cases, the best performing model was plBART, with one exception of AA *body+lead2title* task in which polT-large outperformed the others. Usually, some version of polT (small, base, or large) was the second after plBART. Considering the mean ROUGE averaged over all summarization tasks, plBART is the winner with an average ROUGE 0.293 ± 0.003 and the second is polT-small with the result 0.255 ± 0.062. polT models are always better than mT5 and papuGaPT2. We hypothesize that the pre-training procedure that makes BART biased towards copying the source is a good heuristic in the news summarization task. Furthermore, we observed that large architectures are unstable, and their performance degrades.

Baseline To further evaluate the results, we computed a baseline: copying 3 leading sentences from source as candidate target summarization, following (Gliwa et al., 2019; See et al., 2017). Then we computed a second baseline *adaptive n*: copying n leading sentences from source that best match the average target length determined on the train set. On average $n = 11$ for PSC tasks, $n = 3$ for AA *body2lead* and $n = 1$ for AA *body+lead2title*. Adaptive baseline was much better than fixed $n = 3$.

The best model was usually 2-3pp above the adaptive baseline except for the PSC abstract task (0.1pp below the baseline) and AA *body+lead2title* (32pp above the baseline). However, not all the models outperformed the baseline. PSC tasks have strong baselines. In the news domain, the essential information is at the beginning of the article.

Upper bound estimates PSC dataset contains multiple summaries by different authors per source text. We evaluate human performance as ROUGE metrics of different summaries for the same source text. This upper bound includes both variability in the content and the length of the summary. The best models are halfway between the baseline and upper bound for the PSC whole and PSC extract task. Interestingly, for PSC abstract, all: the baseline, the best model and the upper bound are roughly the same. This shows that ROUGE is not a perfect metric, in particular for abstractive summaries.

3.2.4 Discussion

The PSC dataset contains summaries of different lengths that may degrade performance since the model may be uncertain about the length of the summary. One idea to lift that ambiguity is to add a prefix that would specify the target summary length. Furthermore, different source contexts seen by each model made interpretation of the results more challenging. It was not clear whether the drop in metrics was due to model performance or less information. If the source was trimmed down significantly and the target was not affected, then the whole task turns into generation rather than summarization. On the other hand, copy baselines performed very well on PSC, and indeed, every model saw the leading sentences from the source during training. Almost none of them, except plBART, were able to learn valuable summaries.

After exploration, we found out that, on average, summaries generated by encoder-decoder architectures (polT, mT5, plBART) are shorter than targets on PSC and AA tasks. The papuGaPT2 model generated summaries comparable in length to targets. We also spotted some anomalies. The mT5-large on PSC whole task generated summaries of maximal length, much longer than the target. The issue was not present on PSC abstract task. Furthermore, the PSC abstract was particularly difficult for polT-base which generated extremely short summaries (samples can be found in Appendix G, H, I).

3.3 Question answering

Question answering (QA) systems enable users to automatically obtain accurate answers to questions asked in natural language. We distinguish reading comprehension QA, in which the system, apart from the question, also receives a passage, which may contain the correct answer, and open-domain QA, in which the system receives only the question itself and answers only using the previously collected knowledge.

3.3.1 Datasets

There is no standard dataset for training question answering system for the Polish language, so we

[14] We used native Python implementation of ROUGE score https://github.com/google-research/google-research/tree/master/ROUGE and replaced default tokenizer with nltk http://www.nltk.org/ punkt word tokenizer for Polish language.

Model		ROUGE AVG	AA body2lead	AA body+lead2title	PSC whole	PSC extract	PSC abstract
Baselines	lead n=3 source sentences	17.0	**12.4**	6.8	22.0	23.4	20.3
	lead n (adaptive) source sentences	**21.6**	**12.4**	**7.9**	**30.3**	**31.7**	**25.6**
Upper bounds	human performance	-	-	-	34.3	39.0	25.4
Small models	mT5	23.3 ± 0.4	13.0	34.2	23.3	25.0	21.2
	polT	25.5 ± 6.2	14.3	36.3	30.5	23.2	23.2
Base models	papuGaPT2	15.0 ± 0.2	12.1	14.0	16.6	17.5	14.8
	mT5	20.2 ± 1.3	14.1	36.1	20.6	21.2	8.8
	plBART	**29.3 ± 0.3**	**15.6**	38.3	**32.6**	**34.3**	**25.5**
	polT	23.1 ± 3.3	14.9	38.9	25.2	24.7	11.7
Large models	mT5	15.3 ± 1.1	10.9	33.5	12.0	10.9	9.1
	polT	18.9 ± 1.8	12.1	**39.4**	17.5	15.1	10.6

Table 3: Evaluation results on all summarization tasks (test split, mean values across 3 runs). For simplicity only arithmetic mean ROUGE (f-measure) is reported, ROUGE-1, ROUGE-2 and ROUGE-L are in Appendix D (Tables 10-14). The best scores are in bold, results above the baseline are underlined. ROUGE AVG is arithmetic mean of mean ROUGE of all tasks per model.

combined several resources to create a QA task specifically for this evaluation.

MKQA Multi-lingual Knowledge Questions & Answers (MKQA) (Longpre et al., 2020) contains 10,000 queries sampled from the Google Natural Questions dataset (Kwiatkowski et al., 2019), manually translated from English into 26 typologically diverse languages, including Polish.

After manually inspecting a sample of the data, we removed many of the questions due to specific domain (TV series, movies, etc.), open-endedness ("why?" and "how?" questions), lacking the answer or the answer depending on when question is asked (e.g. "Who won World Cup this year?"). After filtration, we left 1875 useful questions.

Jeden z Dziesięciu is a Polish game show broadcast on Polish Television. The participants answer the host's questions from various domains. We gathered 1004 question-answer pairs[15].

Poleval is an evaluation campaign for NLP tools for Polish. The 2021 edition contains the question answering task[16] with a dataset of 6000 questions and answers. We used the validation set (1000 questions) for training and the test-A set (2500 questions) for test. The test-B set was not released at the time of conducting the experiments.

Final dataset dataset combines the aforementioned datasets, which resulted in the training set of 3879 questions and the test set of 2500 questions.

3.3.2 Tasks

In the task of open-domain question answering we aim to assess the factual knowledge stored in pre-trained models. We expect that the evaluated models already have the required knowledge to answer the question and we use fine-tuning procedure only to guide the models on how to generate the answer (Prager, 2006; Roberts et al., 2020).

The next question answering task is similar to the well-known reading comprehension task: besides the question, the model additionally takes a passage of text which may contain the answer (Zeng et al., 2020). Since none of the used datasets contains

[15] We parsed http://tvturnieje.blogspot.com/p/jeden-z-dziesieciu.html

[16] https://github.com/poleval/2021-question-answering

such passages, we retrieved them on our own in the following way.

First, we manually annotated 10k question-passage pairs with an information if the passage contains the correct answer. The source of candidates for positives pairs were severalfold. We started with a simple Bag-of-Words retriever as well as a more sophisticated model such as Universal Sentence Encoder (Yang et al., 2019). Next, we train a neural retriever based on HerBERT-base model (Mroczkowski et al., 2021) and annotated the retrieved passages.

Overall out of 10k annotated pairs, there were 2215 matching passages for 1347 unique questions[17]. We combined them with "Czy wiesz?" dataset (Marcinczuk et al., 2013; Rybak et al., 2020) to train a neural retriever based on HerBERT-base model.

Finally, to prepare a dataset for the task, we used this retriever to find the top 10 best Wikipedia passages for all questions in our dataset. The goal of the task was to generate the answer based on a question and retrieved passages.

3.3.3 Experiments

Using the aforementioned training set, we finetuned the following models: polT-base, polT-wiki[18], polT-large, mT5-base, plBART, papuGaPT2 and T5 with random weights (T5-random) as a reference point. We trained the models for 50 epochs with Adam (Kingma and Ba, 2014) optimizer except for polT-large, which we trained for 30 epochs. We performed the best learning rate search for each model in the set of {1e-2,5e-3, 1e-3, 5e-4} with 1 seed value. We used a single NVIDIA A100 GPU for all the models. The best results for each model are presented in Table 4. In both tasks the best model was polT-large. In base size polT-wiki performs the best in both tasks.

Models trained on open-ended task learned mainly to answer yes/no questions. In general, they were able to generate a reasonable and fluent response but mostly incorrect. Models also tend to overfit to the training examples from the similar domain (e.g. it answers "Montmartre" for the question of "the highest peak of the Beskids range"). Finetuning the model on Wikipedia as the knowledge database improved the results on both tasks by 1.0pp and 0.3pp, respectively. However, directly providing retrieved passages improves the F1 score by a magnitude of 20pp. Thus, it is evident that there is still plenty of room to improve text-to-text models from a knowledge database perspective.

In the passages subtask we encountered a similar trimming issue as in summarization task 3.2.2. Because of papuGaPT2 input size limitation we couldn't pass the entire passage to the model. That's why we repeated polT-wiki and plBART training using 1024 input size setting. The results for polT and plBART were not significantly different than those with not size-limited input, what confirms their superiority over decoder-only architecture in this task.

Task		open	passages
Base models	polT	20.9	42.8
	polT-wiki	**21.9**	**43.1**
	papuGaPT2	18.5	24.0
	mT5	17.2	39.8
	plBART	17.4	37.1
	T5-random	10.9	8.0
Large models	polT	**26.5**	**51.3**

Table 4: Accuracy scores for open questions and questions followed by the generated passages evaluated using Poleval *testA* set.

3.4 Machine translation

Machine translation is one of the most popular applications of text-to-text models. Therefore, we finetuned polT on parallel corpora consisting of pairs of English and Polish sentences and evaluate performance in both en→pl and pl→en directions.

3.4.1 Datasets

We used datasets collected for the news translation competition organized as part of the 5th Conference on Machine Translation[19] (WMT20). Specifically, we used 4 out of 5 parallel corpora available to the competition's participants: EuroParl v10, TildeRapid, WikiTitles v2, and ParaCrawl v5.1. The corpora are described by Barrault et al. (2020). We did not use the 5th dataset, WikiMatrix, which is known to introduce more noise than useful knowledge (Caswell et al., 2021). To all 4 corpora, we applied similar filtering as Jónsson et al. (2020).

[17] We release the dataset at `anonymized`

[18] polT-base additionally fine-tuned using Wikipedia corpus

[19] https://www.statmt.org/wmt20/translation-task.html

3.4.2 Experiments

We evaluated the models on the WMT20 development set using the BLEU score (Papineni et al., 2002). For generation, we used beam search with 5 beams and a maximal sequence length of 100. Input sentences were also limited to 100 tokens. We compared polT, mT5, plBART, and papuGaPT2. The models were trained in both directions at the same time, with examples fed to the models alternately. Since polT tokenizer was trained on Polish-only data, we trained a separate tokenizer for WMT20 data. It was a unigram language model (Kudo, 2018) with a vocabulary of 32k tokens based on our training corpora with the same filtering. In order to initialize embeddings of tokens from the WMT20 tokenizer that are not present in the polT tokenizer, we applied a technique based on Arkhipov et al. (2019, section 3). To have a fair comparison with mT5, plBART and papuGaPT2, we applied the same conversion to those models. To distinguish between en→pl and pl→en directions, we prepended source sentences with either `<2en>` or `<2pl>` special tokens.

For training we used Adam optimizer (Kingma and Ba, 2014) with learning rate 1e-1 for small models and 1e-2 for base and large ones, gradient accumulation of 8 batches, 10k steps of a linear warm-up, and inverse square root learning rate schedule (Raffel et al., 2020, section 3.1.2). We used different batch sizes for different model sizes to accommodate as many sentences in a batch as possible. Specific batch sizes are listed in Appendix C. We used a single NVIDIA A100 GPU for all the models. Small and base models were trained for 5 epochs and large for 2.

3.4.3 Results

The results are presented in Table 5. polT generates translations superior to mT5 in direction en→pl but not in pl→en. This may be due to the fact that polT was trained on Polish-only data and, therefore, the ability to generate English text deteriorates. polT also outperforms plBART and papuGaPT2. We repeated polT-base and plBART experiments for 3 different seeds. Variance of the polT results was up to 0.1 BLEU, while the variance of the plBART results was slightly higher, up to 0.5 BLEU.

It should be stated that the results reported for polT are not up to the level of the WMT20 winning team's results (Krubiński et al., 2020). However, polT-large performance in the en→pl direction is slightly better than their bi-directional baseline. In this work, we did not attempt to achieve state-of-the-art on those datasets, but to compare polT with similar architectures in similar settings.

Model		Vocab	en→pl	pl→en
Small models	mT5	mT5	20.5	25.0
	polT	polT	20.3	24.7
	mT5	wmt20	<u>20.8</u>	24.8
	polT		<u>20.8</u>	<u>25.4</u>
Base models	mT5	mT5	22.4	<u>27.0</u>
	polT	polT	<u>23.2</u>	26.7
	mT5	wmt20	22.5	26.7
	polT		22.7.	26.8
	plBART	plBART	21.2	25.4
		wmt20	21.6	26.3
	papu-GaPT2	papuGaPT2	21.2	25.5
		wmt20	22.0	26.2
Large models	mT5	wmt20	24.8	**<u>29.0</u>**
	polT		**<u>25.5</u>**	28.9

Table 5: BLEU scores for the WMT20 devset. Best results in each category (small, base and large) are underlined. Best overall results are shown in bold.

4 Conclusion

In this work, we introduced a novel text-to-text model for Polish, polT. It outperforms mT5 on the KLEJ benchmark, summarization, en-pl machine translation, and question answering. It is better than plBART and papuGaPT2, except for summarization, where plBART is the best. We should state that the overall performance of plBART is impressive, considering that it has almost twice as few parameters as polT-base. We observed that the larger the model, the better the results (except the summarization), and encoder-decoder architectures are better than decoder only. We efficiently pretrained polT by initializing the weights from mT5 checkpoints with no exhaustive training. Lastly, since summarization and question answering lack Polish language benchmark datasets, we described their construction and released them publicly.

References

Mikhail Arkhipov, Maria Trofimova, Yuri Kuratov, and Alexey Sorokin. 2019. Tuning multilingual trans-

formers for language-specific named entity recognition. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 89–93, Florence, Italy. Association for Computational Linguistics.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Diedre Carmo, Marcos Piau, Israel Campiotti, Rodrigo Nogueira, and Roberto Lotufo. 2020. Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. *arXiv preprint arXiv:2008.09144*.

Isaac Caswell, Julia Kreutzer, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. 2021. Quality at a glance: An audit of web-crawled multilingual datasets. *arXiv preprint arXiv:2103.12028*.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pages 70–79, Hong Kong, China. Association for Computational Linguistics.

Zolkepli Husein. 2018. Malaya, natural-language-toolkit library for bahasa malaysia, powered by deep learning tensorflow. `https://github.com/huseinzol05/malaya`.

Haukur Páll Jónsson, Haukur Barri Símonarson, Vésteinn Snæbjarnarson, Steinþór Steingrímsson, and Hrafn Loftsson. 2020. Experimenting with different machine translation models in medium-resource settings. In *Text, Speech, and Dialogue - 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8-11, 2020, Proceedings*, volume 12284 of *Lecture Notes in Computer Science*, pages 95–103. Springer.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Mateusz Krubiński, Marcin Chochowski, Bartłomiej Boczek, Mikołaj Koszowski, Adam Dobrowolski, Marcin Szymański, and Paweł Przybysz. 2020. Samsung R&D institute Poland submission to WMT20 news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 181–190, Online. Association for Computational Linguistics.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoit Crabbé, Laurent Besacier, and Didier Schwab. 2020. FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France. European Language Resources Association.

Mike Lewis, 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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yinhan Liu, 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*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2020. Mkqa: A linguistically diverse benchmark for multilingual open domain question answering.

Michał Marcinczuk, Marcin Ptak, Adam Radziszewski, and Maciej Piasecki. 2013. Open dataset for development of polish question answering systems. In *Proceedings of the 6th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, Wydawnictwo Poznanskie, Fundacja Uniwersytetu im. Adama Mickiewicza*.

Louis Martin, Benjamin Muller, Pedro Javier Ortiz Suárez, Yoann Dupont, Laurent Romary, Éric de la Clergerie, Djamé Seddah, and Benoît Sagot. 2020. CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Robert Mroczkowski, Piotr Rybak, Alina Wróblewska, and Ireneusz Gawlik. 2021. HerBERT: Efficiently pretrained transformer-based language model for Polish. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 1–10, Kiyv, Ukraine. Association for Computational Linguistics.

El Moatez Billah Nagoudi, Wei-Rui Chen, Muhammad Abdul-Mageed, and Hasan Cavusoglu. 2021a. IndT5: A text-to-text transformer for 10 indigenous languages. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 265–271, Online. Association for Computational Linguistics.

El Moatez Billah Nagoudi, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. 2021b. Arat5: Text-to-text transformers for arabic language understanding and generation.

Maciej Ogrodniczuk and Mateusz Kopeć. 2014. The Polish summaries corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3712–3715, Reykjavik, Iceland. European Language Resources Association (ELRA).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

John Prager. 2006. Open-domain question: Answering. *Found. Trends Inf. Retr.*, 1(2):91–231.

Adam Przepiórkowski, Mirosław Bańko, Rafał L Górski, Barbara Lewandowska-Tomaszczyk, Marek Łaziński, and Piotr Pęzik. 2011. National corpus of polish. In *Proceedings of the 5th language & technology conference: Human language technologies as a challenge for computer science and linguistics*, pages 259–263.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Colin Raffel, 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*, 21:1–67.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Piotr Rybak, Robert Mroczkowski, Janusz Tracz, and Ireneusz Gawlik. 2020. KLEJ: Comprehensive benchmark for Polish language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1191–1201, Online. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. 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)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Noam Shazeer, Youlong Cheng, Niki Parmar, Dustin Tran, Ashish Vaswani, Penporn Koanantakool, Peter Hawkins, HyoukJoong Lee, Mingsheng Hong, Cliff Young, et al. 2018. Mesh-tensorflow: deep learning for supercomputers. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 10435–10444.

Noam Shazeer and Mitchell Stern. 2018. Adafactor: Adaptive learning rates with sublinear memory cost.

Antti Virtanen, Jenna Kanerva, Rami Ilo, Jouni Luoma, Juhani Luotolahti, Tapio Salakoski, Filip Ginter, and Sampo Pyysalo. 2019. Multilingual is not enough: Bert for finnish.

Linting Xue, 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. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, et al. 2019. Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint arXiv:1907.04307*.

Changchang Zeng, Shaobo Li, Qin Li, Jie Hu, and Jianjun Hu. 2020. A survey on machine reading comprehension: Tasks, evaluation metrics and benchmark datasets.

A Converting KLEJ benchmark to text-to-text format

A.1 Task formulation

We transformed KLEJ tasks in a text-to-text format where each input sentence (one or two in KLEJ tasks) is given a prompt describing its role in the task semantics. Additionally, a task identification token could prefix whole input, but in all experiments in the paper, we did not use such prompt because we were training models for one task at a time. Finally, labels were converted into semantically significant text tokens. Input features and task labels are given in the Table 6.

A.2 Evaluation

At test time model is given input formatted as described above, and the goal is to generate a label. A sequence of target tokens is generated until EOS token generation or when it reaches max target length. We calculate the max target length per task, and it corresponds to the longest label after the tokenization process. The generated sequence of tokens is converted into text. We count for a positive when it matches precisely the corresponding label.

A.3 Learning rate tuning

We performed the best learning rate search for each model in the set of {2e-5, 4e-5, 6e-5, 8e-5, 1e-4, 2e-4, 4e-4, 6e-4, 8e-4, 1e-3}. For learning rate tuning, we used tasks with the validation set provided. The designated learning rate value was used for all test runs on all KLEJ tasks.

B Summarization task

B.1 Allegro Reviews (AA)

The raw dataset is a collection of articles scrapped from `https://allegro.pl/artykuly` website. Each example contains title, lead, body, information about category tree, and other metadata. In a pre-processing stage, we removed markdown formatting and normalized white spaces. This dataset serves as a good benchmark for highly abstractive summarization/generation tasks since every article contains a title or lead that can be viewed as a loose abstract of the article. Moreover, every article is written by a professional editor, and they are relatively short hence fit almost entirely into 512 token context.

We constructed 2 versions of the abstractive summarization task, which affect the summary ratio. In *body2lead* task, we use a body of the article as a source and lead as a generation target. In *body+lead2title* we use concatenated lead and body as source (full article) and generate title as a target.

For evaluation purposes, we divided the whole data source into 80% train set and 20% test set in a stratified way using top-level category in Allegro category tree as a label. Then, we saved 10% of the train set for validation (also stratified split) and hyperparameter tuning. Table 7 contains summary of average source/target length and size of each split.

B.2 Polish Summaries Corpus (PSC)

Original dataset (Ogrodniczuk and Kopeć, 2014) contains extractive or abstractive summaries and other metadata for news articles prepared by annotators[20]. Each article has a few corresponding summaries that depend on the annotator, summary ratio (5, 10, or 20% of the original text), and type (extractive or abstractive). Due to the small size of the entire dataset, we included all summaries in the final dataset. In this way, the source contains duplicates and targets near-duplicates. Moreover,

[20] We used publicly available version `https://huggingface.co/datasets/polsum`

Task	Prefix 1	Prefix 2	Labels
NKJP-NER	zdanie	-	geograficzna, brak, organizacja, osoba, miejsce, czas
CBD	zdanie	-	neutralna, przemoc
Czy wiesz?	pytanie	odpowiedź	fałsz, prawda
PolEmo2.0	zdanie	-	niejednoznaczny, negatywny, pozytywny, neutralny
AR	zdanie	-	1.0, 2.0, 3.0, 4.0, 5.0
PSC	streszczenie 1	streszczenie 2	nie_parafraza, parafraza
CDSC-E	zdanie 1	zdanie 2	neutralny, wynikanie, sprzeczność

Table 6: Prompts and labels used in text-to-text formulation of KLEJ benchmark tasks. In case of encoder-decoder architectures, source was prepared as `<Prefix 1>`: text 1 `<Prefix 2>`: text 2. For decoder architecture we additionally append `[SEP]` token at the end.

our train set contains target summaries of different ratios that may affect learning and prediction.

We constructed 3 versions of the summarization task. The *extract* subset contains only extractive summaries, the *abstract* subset contains only abstractive summaries and *whole* contains both of them.

For evaluation purposes, we divided the whole data source into 80% train set and 20% test set in a stratified way using summary ratio, summary type, and article category as a label. Then, we saved 10% of the train set for validation (also stratified split) and hyperparameter tuning. Table 7 contains summary of average source/target length and size of each split.

Task	AA body2lead	AA body+lead2title	PSC whole	PSC extract	PSC abstract
Domain	E-commerce		News articles		
Average length (characters)					
Source	3.9k	4.1k	10.6k		
Target	0.3k	0.05k	1.3k		
Number of examples					
Train	24.4k		0.8k	6.1k	1.7k
Dev	2.7k		0.9k	0.7k	0.2k
Test	6.8k		2.2k	1.7k	0.5k

Table 7: Overview of summarization tasks with total number of examples for each split and average length (in characters) of source and target

B.3 Trimming inputs

The BART and GPT-2 architectures use fixed positional encoding, which is limited to 1024 tokens by default. On the other hand, the T5 model uses relative positional encoding and can process sequences of any length in principle. In our case, we are bounded by memory requirements. Thus, we fixed 1024 tokens as the maximal sequence length for all models. Even then, large architectures required further trimming of input tensors due to memory errors. Consequently, we were not able to fit the whole target and source texts into the model inputs.

Since each model uses a different tokenizer, the context used during training and prediction was different for individual models. The effect is more pronounced for source texts because they are much longer than targets. We were always able to input 98-100% of target tokens into the model and assumed that target trimming did not substantially affect the results. Usually, only a small number of outliers (extremely long targets) were trimmed.

On the other hand, source trimming could have affected the models. The final number of tokens to which the source was trimmed and the percentage of the source tokens seen by the model are shown in Table 8. The PSC task was affected the most. For example, mT5-large saw on average only 25% of the source. Similarly, the papuGaPT2 saw on average only 26% of the source. Less affected were plBART and polT models, which saw above 50% of the source. Most importantly, none of the models ever saw 100% of source texts during training for PSC tasks.

B.4 Learning rate tuning

On AA task, we performed best learning rate search in the range {1e-5, 2e-5, 4e-5, 6e-5, 8e-5, 1e-4, 2e-4, 4e-4, 6e-4, 8e-4, 1e-3}. On PSC task, we per-

Task		PSC	AA body2lead	AA body+lead2title
Small models	mT5	33%	85%	80%
	polT	51%	96%	95%
Base models	papuGaPT2	26% (512)	92% (767)	96% (959)
	mT5	33%	85%	80%
	plBART	52%	96%	95%
	polT	51%	96%	95%
Large models	mT5	25% (768)	85%	80%
	polT	51%	96%	95%

Table 8: Average percentage of tokens from source text that fit into model. Differences are due to a tokenizer, model capacity, and memory limitations. By default, source text was trimmed to 1024 tokens, and the exceptions are indicated (in parentheses) in the table.

formed learning rate search for each model on PSC whole subset and used the same learning rates for PSC extract and PSC abstract subsets as well. The search was geometrical, with the start point at 4e-4, i.e. at first subset {2e-4, 4e-4, 8e-4} was examined, each learning rate with 3 seed runs. If the middle point was better than the two other, the search completed, otherwise either {1e-4, 2-e4, 4e-4} or {4e-4, 8e-4, 1.6e-3} learning rates were examined during the next step depending on the trend. As the result, learning rates from 5e-5 to 2.56e-2 were selected for different models. The only exception was mT5-large model which was unstable and required finer granularity {2.5e-5, 5e-5, 7.5e-5 } was used at the end of the search. Our procedure of learning rate tuning was motivated by budget constraints.

C Machine translation batch sizes

All translation experiments were run on a single NVIDIA A100 GPU with 40 GB of memory. To utilize all the available memory we used different batch sizes for different model architectures. Specific batch sizes are shown in Table 9.

Model		Vocab	Batch size
Small models	mT5	mT5	25
	polT	polT	50
	mT5 polT	wmt20	50
Base models	mT5	mT5	10
	polT	polT	25
	mT5 polT	wmt20	25
	plBART	plBART wmt20	40
	papu-GaPT2	papuGaPT2 wmt20	25
Large models	mT5 polT	wmt20	10

Table 9: Machine translation batch sizes

D Summarization ROUGE scores

Model		ROUGE-1	ROUGE-2	ROUGE-L	ROUGE
Baselines	lead n=3 source sentences	**20.3**	**2.8**	**14.1**	**12.4**
	lead n (adaptive, avg n=3) source sentences	20.2	**2.8**	**14.1**	**12.4**
Small models	mT5	19.7	4.1	15.2	13.0 ± 0.1
	polT	21.5	4.9	16.5	14.3 ± 0.1
Base models	papuGaPT2	20.2	1.6	12.6	12.1 ± 0.1
	mT5	21.4	4.6	16.3	14.1 ± 0.2
	polT	22.3	5.2	17.1	14.9 ± 0.1
	plBART	**24.0**	**5.1**	**17.7**	**15.6 ± 0.1**
Large models	mT5	16.7	2.7	13.4	10.9 ± 2.4
	polT	19.2	2.7	14.4	12.1 ± 2.6

Table 10: Evaluation results on Allegro Articles body2lead task (test split, mean values across 3 runs). The best scores are in bold, results above the baseline are underlined. The ROUGE score reported in the last column is arithmetic mean of ROUGE-1, ROUGE-2 and ROUGE-L scores. All reported ROUGE scores are f-measures.

Model		ROUGE-1	ROUGE-2	ROUGE-L	ROUGE
Baselines	lead n=3 source sentences	9.3	2.9	8.1	6.8
	lead n (adaptive, avg n=1) source sentences	**10.4**	**3.6**	**9.7**	**7.9**
Small models	mT5	39.3	25.1	38.2	34.2 ± 0.3
	polT	41.4	27.2	40.1	36.3 ± 0.1
Base models	papuGaPT2	18.8	5.2	18.0	14.0 ± 0.1
	mT5	41.3	27.0	40.1	36.1 ± 1.2
	plBART	44.0	29.6	41.3	38.3 ± 0.1
	plBART†	44.0	29.7	42.4	38.7 ± 0.5
	polT	44.1	29.7	42.8	38.9 ± 0.3
	polT†	43.9	29.4	42.6	38.7 ± 0.2
Large models	mT5	38.4	24.7	37.3	33.5 ± 1.8
	polT	**44.6**	**30.2**	**43.3**	**39.4 ± 0.4**

Table 11: Evaluation results on Allegro Articles body+lead2title task (test split, mean values across 3 runs). The best scores are in bold, results above the baseline are underlined. The ROUGE score reported in the last column is arithmetic mean of ROUGE-1, ROUGE-2 and ROUGE-L scores. All reported ROUGE scores are f-measures. (†) models were trained with the same input and target truncation as papuGaPT2.

Model		ROUGE-1	ROUGE-2	ROUGE-L	ROUGE
Baselines	lead n=3 source sentences	28.2	16.5	21.3	22.0
	lead n (adaptive, avg n=11) source sentences	**39.0**	**23.8**	**28.3**	**30.3**
Upper bounds	human performance	41.8	27.4	33.8	34.3
Small models	mT5	29.4	17.0	23.5	23.3 ± 0.5
	polT	36.7	24.3	30.4	30.5 ± 0.1
Base models	papuGaPT2	26.9	7.6	15.5	16.7 ± 0.1
	mT5	27.4	14.1	20.3	20.6 ± 0.2
	plBART	**39.0**	**26.3**	**32.4**	**32.6 ± 0.1**
	polT	32.9	18.1	24.5	25.2 ± 0.5
Large models	mT5	20.6	3.7	11.5	12.0 ± 0.3
	polT	23.1	12.0	17.4	17.5 ± 0.6

Table 12: Evaluation results on Polish Summary Corpus (whole) task (test split, mean values across 3 runs). The best scores are in bold, results above the baseline are underlined. The ROUGE score reported in the last column is arithmetic mean of ROUGE-1, ROUGE-2 and ROUGE-L scores. All reported ROUGE scores are f-measures.

Model		ROUGE-1	ROUGE-2	ROUGE-L	ROUGE
Baselines	lead n=3 source sentences	29.5	18.3	22.6	23.4
	lead n (adaptive, avg n=11) source sentences	**40.0**	**25.5**	**29.7**	**31.7**
Upper bounds	human performance	45.1	33.3	38.8	39.0
Small models	mT5	30.4	19.3	25.3	25.0 ± 0.2
	polT	27.7	18.3	23.6	23.2 ± 13.8
Base models	papuGaPT2	26.9	8.5	15.9	17.1 ± 0.2
	mT5	27.6	15.0	20.9	21.2 ± 0.5
	plBART	**39.9**	**28.6**	**34.4**	**34.3 ± 0.1**
	polT	31.9	18.0	24.1	24.7 ± 0.5
Large models	mT5	19.3	2.6	10.9	10.9 ± 1.4
	polT	22.4	7.3	15.6	15.1 ± 2.6

Table 13: Evaluation results on Polish Summary Corpus (extract) task (test split, mean values across 3 runs). The best scores are in bold, results above the baseline are underlined. The ROUGE score reported in the last column is arithmetic mean of ROUGE-1, ROUGE-2 and ROUGE-L scores. All reported ROUGE scores are f-measures.

Model		ROUGE-1	ROUGE-2	ROUGE-L	ROUGE
Baselines	lead n=3 source sentences	0.271	0.142	0.197	0.203
	lead n (adaptive, avg n=11) source sentences	**0.354**	**0.178**	**0.236**	**0.256**
Upper bounds	human performance	0.346	0.167	0.250	0.254
Small models	mT5	0.280	0.149	0.207	0.212 ± 0.005
	polT	0.304	0.165	0.227	0.232 ± 0.019
Base models	papuGaPT2	0.260	0.057	0.139	0.152 ± 0.002
	mT5	0.120	0.050	0.095	0.088 ± 0.025
	plBART	**0.335**	**0.177**	**0.254**	$\mathbf{0.255 \pm 0.007}$
	polT	0.156	0.077	0.120	0.117 ± 0.073
Large models	mT5	0.161	0.019	0.094	0.091 ± 0.008
	polT	0.156	0.051	0.112	0.106 ± 0.011

Table 14: Evaluation results on Polish Summary Corpus (abstract) task (test split, mean values across 3 runs). The best scores are in bold, results above the baseline are underlined. The ROUGE score reported in the last column is arithmetic mean of ROUGE-1, ROUGE-2 and ROUGE-L scores. All reported ROUGE scores are f-measures.

E Example predictions on Allegro Reviews *body+lead2title*

1. **Gold Summary:** Speedminton – badminton na wietrzne dni
   polT-base: Jak zacząć grać w speedminton?
   plBART: Speedminton – sport dla każdego
   papuGaPT2: Squa – jak zacząć?
   mT5-base: Jak zacząć grać w speedmintona?
   polT-large: Speedminton – sport dla aktywnych

2. **Gold Summary:** Urządzamy wnętrze po skandynawsku
   polT-base: Jak urządzić mieszkanie w stylu skandynawskim?
   plBART: Jak urządzić mieszkanie w stylu skandynawskim?
   papuGaPT2: Skandynawski styl w wersji budżetowej – jak go urządzić?
   mT5-base: Jak urządzić wnętrze w stylu skandynawskim?
   polT-large: Jak urządzić mieszkanie w stylu skandynawskim?

3. **Gold Summary:** Czy warto robić domowe peelingi – poradnik
   polT-base: Najlepsze peelingi do domu
   plBART: Najlepsze peelingi do twarzy
   papuGaPT2: Peeslingi – najlepszy sposób na przesuszoną skórę
   mT5-base: Jak zrobić peelingi w domu?
   polT-large: Jak zrobić peeling w domu?

4. **Gold Summary:** Szerokie spodnie – must have sezonu wiosna/lato 2015
   polT-base: Wiosenne trendy – szerokie spodnie
   plBART: Modne szerokie spodnie – jak je nosić?
   papuGaPT2: Wiosenne stylizacje ze schullami w roli głównej
   mT5-base: Wiosenny trend – szerokie spodnie
   polT-large: Wiosenne spodnie w stylu lat 70.

5. **Gold Summary:** Przygotuj się na lodowisko
   polT-base: Jak rozpocząć sezon na łyżwy?
   plBART: Jak zacząć przygodę z łyżwami?
   papuGaPT2: Jakie ubrania na sezon kąpielowy?
   mT5-base" Jak zacząć sezon na łyżwy?
   polT-large: Jak przygotować się do sezonu na łyżwy?

6. **Gold Summary:** Wózek widłowy niczym limuzyna sterowana iPadem
   polT-base: Wózki widłowe zwiększające możliwości swojego działania
   plBART: Wózki widłowe – zwiększamy wydajność i oszczędzamy paliwo
   papuGaPT2: Nowoczesny wózek widłowy – dlaczego jest tak ważny?
   mT5-base: Czym są nowoczesne technologie w wózkach widłowych?
   polT-large: Automatyzacja wózków widłowych

F Example predictions on Allegro Reviews *body2lead*

1. **Gold Summary:** Każdy czuje respekt przed królową nauk. Matematyka spędza sen z powiek niejednemu uczniowi. Są jednak sposoby zachęcenia dziecka do nauki matematyki i sprawienia, że lepiej ją zrozumie i polubi.

   polT-base: Matematyka to jedna z najprzyjemniejszych dziedzin nauki. Uczymy się ją w szkole i w domu. Nauka przez zabawę może być świetną rozrywką dla całej rodziny. Jakie gadżety mogą pomóc dziecku w nauce liczenia?

   plBART: Nauka matematyki nie musi być trudna. Wystarczy odrobina chęci i kilka akcesoriów, które pomogą dziecku w odkrywaniu tajemnic matematyki.

   papuGaPT2: Nauka poprzez zabawę to jeden z najlepszych sposobów na rozwijanie zdolności poznawczych i umiejętności logicznego myślenia. Podpowiadamy, co warto zabrać ze sobą do szkoły i na co zwrócić uwagę przy wyborze odpowiednich akcesoriów.

   mT5-base: Nauka matematyki to doskonały sposób na spędzenie wolnego czasu w gronie najbliższych. Dzieci chętnie sięgają po pomoce naukowe, które zachęcą do nauki przez zabawę i doświadczenie.

   polT-large: Matematyka nie musi być nudna. Nauka liczenia może być wspaniałą zabawą, która wspomoże rozwój dziecka. Podpowiadamy, jak zachęcić dziecko do nauki matematyki.

2. **Gold Summary:** Ciepłe, letnie dni aż się proszą o aktywną zabawę na dworze. Dlatego jeśli planujesz przyjęcie dla dziecka i jego przyjaciół, koniecznie postaw na garden party. Tego typu imprezy są nie tylko modne, ale także łatwe w przygotowaniu i posprzątaniu po nich. Jeśli nie masz pomysłu, jak je przygotować, podpowiadamy.

   polT-base: Ozdabianie ogrodu to nie tylko świetna zabawa, ale także świetna rozrywka. Podpowiadamy, jak przygotować ogrodowe przyjęcie dla dziecka.

   plBART: Dzieci uwielbiają spędzać czas na świeżym powietrzu. To doskonała okazja do zorganizowania im urodzinowego przyjęcia, które zachwyci ich nie tylko wyglądem, ale także kreatywnością i zapewni wiele pozytywnych emocji. Podpowiadamy, jak przygotować party dla maluchów w cenie do 50 zł.

   papuGaPT2: Organizujesz przyjęcie urodzinowe dla dziecka? Zastanawiasz się, co mu kupić, aby było mu jak najlepiej. Podpowiadamy, na co zwrócić uwagę, wybierając dekoracje do ogrodu. Sprawdź nasze propozycje.

   mT5-base: Ogrodowe przyjęcie to wyjątkowe wydarzenie w życiu każdego dziecka. To świetna okazja do spędzenia czasu na świeżym powietrzu. Warto więc zadbać o to, aby było przyjemne i kolorowe. Podpowiadamy, jakie dekoracje będą najlepsze dla twojego malucha.

   polT-large Ozdabianie ogrodu to nie tylko świetna zabawa, ale także świetna rozrywka. Podpowiadamy, jak przygotować ogrodowe przyjęcie dla dziecka.

G Example predictions on PSC (whole)

1. **Gold Summary:** Ostatnia znaczna podwyżka stóp procentowych NBP oraz gwałtowne przyspieszenie inflacji wywołały krytykę poczynań Rady Polityki Pieniężnej. Rada chciałaby podjąć dyskusję z głosami krytycznymi. nikt nie przewidywał, że w 1999 r. inflacja przewyższy cel inflacyjny NBP.Świadczy to dobitnie o tym, że głównymi czynnikami przyspieszającymi inflację były szoki podażowe na rynku żywności i paliw. Inny komentator stawia z kolei zarzut, że Rada niepotrzebnie zwlekała z decyzją o podwyżce stóp do listopada. Rada listopadową decyzję nie dlatego podjęła w listopadzie, i w takiej skali, że "zaspała" w październiku, ale dlatego, by podjąć ją w takiej właśnie skali w listopadzie. Dla każdego, kto wie, co to jest efektywność mechanizmu transmisji impulsów polityki pieniężnej do gospodarki, decyzja taka jest zrozumiała. Każdy ma prawo twierdzić, że dokonane podniesienie stóp NBP jest za duże. Pamiętać jednak musi, że jego twierdzenie sprowadza się do tezy: skala dokonanej podwyżki stóp procentowych doprowadzi do przestrzelenia celu inflacyjnego w dół. Czy rację mają ci, którzy krytykują teraz Radę, dowiemy się za rok. Przejdźmy do zarzutu zbyt dużej redukcji stóp procentowych 20 stycznia 1999 r. Otóż, redukcja rzeczywiście była zbyt duża bądź niepotrzebna. Problem polega jednak na tym, że o tym wiemy dopiero teraz. skala obniżki stóp procentowych dostosowana była do przewidywanego w 1999 r. przebiegu zjawisk makroekonomicznych. Było to działanie zgodne z zasadą forward looking. Czy Rada miała przesłanki do wcześniejszej zmiany swoich przewidywań? Bardzo mało i późno. o problemach ZUS dowiedzieliśmy się dopiero w lipcu. Gwałtowny wzrost cen żywności rozpoczął się od sierpnia. wzrosły ceny paliw. Z żadnych analiz nie wynikała taka skala wzrostu cen na rynku żywności i paliw. W tym kontekście stawianie Radzie zarzutu, że będąc odpowiedzialną za walkę z inflacją, stara się zrzucić winę na rząd i wydarzenia na rynku międzynarodowym, jest nierzetelne. Nie można winić straży pożarnej za lekkomyślność dziecka z zapałkami, jeśli okaże się sprawna w gaszeniu pożaru. nie można ignorować krótkookresowego wpływu na inflację czynników pozapieniężnych. nie da się na trwałe zmniejszyć inflacji bez budowy sprawnych i konkurencyjnych rynków paliw czy energii.

   mT5-small: Ostatnia znaczna podwyżka stóp procentowych NBP oraz gwałtowne przyspieszenie inflacji wywołały ze strony niektórych ekonomistów, analityków i komentatorów krytykę poczynań Rady Polityki Pieniężnej. niestety, trudno podjąć merytoryczną dyskusję z głosami, szczególnie w sytuacji niepowodzeń.Rada listopadową decyzję podejmowała w całkowitym spokoju, bez żadnej nerwowości.Piędą zwracać uwagę na to, że polityka pieniężna była zbyt pożądanym, zważywszy rolę, jaką odgrywają. W tym kontekście stawianie Radzie zarzutów można było doprowadzić do końca września.

   polT-small: Ostatnia podwyżka stóp procentowych NBP oraz gwałtowne przyspieszenie inflacji wywołały krytykę poczynań Rady Polityki Pieniężnej. Rada listopadową decyzję podejmowała w całkowitym spokoju. Dla każdego, kto wie, co to jest efektywność mechanizmu transmisji impulsów polityki pieniężnej do gospodarki i jakie jest znaczenie oczekiwań w tym procesie, decyzja taka jest zrozumiała. redukcja stóp w styczniu rzeczywiście była zbyt duża bądź w ogóle niepotrzebna.

   papuGaPT2: redukcja stóp w styczniu była zbyt duża bądź w ogóle niepotrzebna. skala krytycznych wniosków była dość trudna, stawiano wiele pytań, do których nie podejmowano ostatecznych decyzji.Rada miała przesłanki do wcześniejszej zmiany swoich przewidywań bardzo późno. informacje o wszystkich wydarzeniach dostępne były kilka tygodni po ich wystąpieniu. niską przewidywalność polityki gospodarczej rządu widać było gołym okiem. decyzja taka jest zrozumiała. redukcja stóp była koniecznym posunięciem, aby zbliżyć się do rynku. jednak Rada przyjmowała wówczas za wiarygodną zapowiedź zaostrzenia polityki fiskalnej oraz przewidywała powolny wzrost eksportu. starano się reanimować kryzys. pojawiła się propozycja, żeby Rada zaczęła mówić jednym głosem. Dziś grono to znacznie się powiększyło. z kolejnych wypowiedzi zaczęli wycofać wniosek o rewizję nadzwyczajną. Rada listopadową decyzję podejmowała w całkowitym spokoju, bez żadnej nerwowości. Dla każdego, kto wie, co to jest efektywność mechanizmu transmisji impulsów polityki do gospodarki i jakie jest znaczenie oczekiwań w tym procesie, decyzja o podwyżce stóp nie jest żadnym donosem. To oznacza duże prawdopodobieństwo żywiołowego narastania deficytu sektora pub-

licznego. Ryzyko destabilizacji rynku jest wysokie. nie wolno zrezygnować z powrotu na przyzwoitą ścieżkę wzrostu ani z poprawienia równowagi zewnętrznej. Trzeba bardziej powściągliwie określić możliwości wycofania się inflacji z polityki monetarnej. jeśli spojrzymy na deprecjację parytetu, okazuje się, ze zestrzelenie nastąpiło nie z inicjatywy samego ROR-u, lecz z chęci ponownego ożywienia poprzez odpartyjnienie poprzez podjęcie działań, których w tej sytuacji jeszcze nie jesteśmy w stanie podjąć.

mT5-base: Nie da się na trwałe zmniejszyć inflację bez budowy sprawnych i konkurencyjnych rynków paliw czy energii Szukanie winnego RYS. MARCIN CHUDZIK BoGUSŁAW GRABOWSKI Ostatnia znaczna podwyżka stóp procentowych NBP oraz gwałtowne przyspieszenie inflacji wywołały ze strony niektórych ekonomistów krytykę poczynań Rady Polityki Pieniężnej. Niestety, trudno podjąć merytoryczną dyskusję z głosami, szczególnie w sytuacji niepowodzeń w ograniczaniu inflacją.

plBART: Ostatnia znaczna podwyżka stóp procentowych NBP oraz gwałtowne przyspieszenie inflacji wywołały ze strony niektórych komentatorów krytykę poczynań Rady Polityki Pieniężnej. Rada listopadową decyzję podejmowała w całkowitym spokoju. Dla każdego, kto wie, co to jest efektywność mechanizmu transmisji impulsów polityki pieniężnej do gospodarki i jakie jest znaczenie oczekiwań w tym procesie, decyzja taka jest w pełni zrozumiała. redukcja stóp w styczniu rzeczywiście była zbyt duża bądź niepotrzebna. o tym wiemy dopiero teraz. stawianie Radzie zarzutu, że będąc odpowiedzialną za walkę z inflacją, stara się zrzucić winę na rząd, jest nierzetelne.

polT-base: Rząd Polityki Pieniężnej w poczuciu odpowiedzialności przywiązuje wielkie znaczenie do przejrzystości swoich działań od początku istnienia. Dlatego chciałaby podjąć dyskusję z głosami krytycznymi, szczególnie w sytuacji niepowodzeń w ograniczaniu inflacji. Znaczna część ostatnich krytycznych wypowiedzi wobec Rady sprowadzała się właściwie do trzech zarzutów: o to, że ostatnia podwyżka stóp procentowych była "spóźnioną, zbyt nerwową i przesadną" reakcją Rady, której "optymizm co do inflacji trwał zbyt długo", i która "jakby próbowała nadrobić zaległości w walce z inflacją".

2. **Gold Summary:** Dwaj słynni bokserzy zawodowi Amerykanin Evander Holyfield i Brytyjczyk Lennox Lewis szykują się do rewanżowego pojedynku. Ich pierwszy pojedynek zakończył się, haniebnym według fachowców, remisem. Wygrany będzie królem boksu zawodowego.Historia boksu jest pełna ekscytujących rewanży, mają je również na swoim koncie Holyfield i Lewis.Obecny faworyt Lewis, jest aktualnym mistrzem świata WBC. Początkowy faworyt Holyfield, ma dwa pasy mistrzowskie WBA i IBF. Jeśli przegra Holyfield przegra tym samym Don King, najsłynniejszego promotora boksu zawodowego.Walkę transmituje Canal Plus.

   mT5-small: W najbliższą sobotę w Las Vegas dojdzie do kolejnego "rewanżu stulecia". Zmierzą się w nim dwaj najwybitniejsi obecnie bokserzy wagi ciężkiej, Amerykanin Evander Holyfield i Brytyjczyk Lennox Lewis. Osiem miesięcy temu, 13 marca, w nowojorskiej Madison Square Garden ich pierwsza walka zakończyła się kontrowersyjnym remisem. Stawką rewanżowego pojedynku będą pasy mistrzowskie trzech najważniejszych organizacji boksu zawodowego: WBA, WBC, IBF.

   polT-small: W najbliższą sobotę w Las Vegas dojdzie do kolejnego "rewanżu stulecia". Zmierzą się w nim dwaj najwybitniejsi obecnie bokserzy wagi ciężkiej, Amerykanin Evander Holyfield i Brytyjczyk Lennox Lewis. Osiem miesięcy temu, 13 marca, w nowojorskiej Madison Square Garden ich pierwsza walka zakończyła się kontrowersyjnym remisem. Stawką rewanżowego pojedynku będą pasy mistrzowskie trzech najważniejszych organizacji boksu zawodowego: WBA, WBC, IBF. Ostatnim posiadaczem trzech pasów w wadze ciężkiej był Riddick Bowe. Od siedmiu lat żaden z pięściarzy nie może się uważać za króla tej najbardziej cenionej kategorii.

   papuGaPT2: W nocy z 13 na 14 listopada w Thomas and Mack Center w "Ready World Press" zmierzą się dwaj znakomici zawodnicy. Ich walkę będzie można oglądać na żywo w Canal Plus od godz. 3:25. Retransmisja w niedzielę będzie miała miejsce już raz trzeci w historii walki, która miała być największym wydarzeniem ostatnich lat. Transmitowany przez Jamesa "Bustera" pojedynek

Sonny'ego Listona z Bobem Fitzsimmonsem w 1899 r. zakończył się dyskwalifikacją jednego z zawodników. Walka okazała się skandalem. Trzy niezależne postępowania wyjaśniające, które prowadzono po tej walce, niewiele dały. Nikomu nie udowodniono winy. \nBoks jest tak stary jak świat. Już Homer opiewał walki na gołe pięści. Wielkim zwolennikiem walk bokserów był rzymski cesarz Kaligula. W 1350 roku, gdy naprzeciw siebie stanęli Jeffries i O'Donnnel, dwóch zawodników zadało przeciwnikowi dwa razy więcej ciosów, niż spalonych. Dwudziestotrzyletni wówczas Buster trafił prawym prostym. Teraz, na dziesięć metrów, już w pierwszej rundzie Floyda Pattersona spotkał się z Joem Frazierem, Muhammadem Alim, Tunneyem i Michaelem Moorerem. O 3.00 rano, kiedy stoczyli już swoje ostatnie rugi, przyszedł czas na jeszcze jedną walkę z "Milionerem", "Drakula" z Jackiem Dempseyem.

mT5-base: W najbliższą sobotę w Las Vegas dojdzie do kolejnego "rewanżu stulecia". Zmierzą się w nim dwaj najwybitniejsi obecnie bokserzy wagi ciężkiej, Amerykanin Evander Holyfield i Brytyjczyk Lennox Lewis. Osiem miesięcy temu, 13 marca, w nowojorskiej Madison Square Garden ich pierwsza walka zakończyła się kontrowersyjnym remisem. Stawką rewanżowego pojedynku będą pasy mistrzowskie trzech najważniejszych organizacji boksu zawodowego: WBA, WBC, IBF.

plBART: W najbliższą sobotę w Las Vegas dojdzie do kolejnego "rewanżu stulecia". Zmierzą się w nim dwaj najwybitniejsi obecnie bokserzy wagi ciężkiej, Amerykanin Evander Holyfield i Brytyjczyk Lennox Lewis. Osiem miesięcy temu, 13 marca, w nowojorskiej Madison Square Garden ich pierwsza walka zakończyła się kontrowersyjnym remisem. Stawką rewanżowego pojedynku będą pasy mistrzowskie trzech najważniejszych organizacji boksu zawodowego: WBA, WBC, IBF. Ostatnim posiadaczem trzech pasów w wadze ciężkiej był Riddick Bowe. Od siedmiu lat żaden z pięściarzy nie może się uważać za króla tej najbardziej cenionej kategorii.

polT-base: W najbliższą sobotę w Las Vegas dojdzie do kolejnego "rewanżu stulecia". Zmierzą się w nim dwaj najwybitniejsi obecnie bokserzy wagi ciężkiej, Amerykanin Evander Holyfield i Brytyjczyk Lennox Lewis. Osiem miesięcy temu, 13 marca, w nowojorskiej Madison Square Garden ich pierwsza walka zakończyła się kontrowersyjnym remisem. Stawką rewanżowego pojedynku będą pasy mistrzowskie trzech najważniejszych organizacji boksu zawodowego: WBA, WBC, IBF. Ostatnim posiadaczem trzech pasów w wadze ciężkiej był Riddick Bowe. Od siedmiu lat żaden z pięściarzy nie może się uważać za króla tej najbardziej cenionej kategorii.

H Example predictions on PSC (extract)

1. **Gold Summary:** Politycy z pierwszych stron gazet szykują się do objęcia stanowisk prezydentów dużych miast. Dla nich gotowi są nawet zrezygnować z mandatów poselskich.\nW najbliższych tygodniach rozstrzygnie się, czy w jesiennych wyborach samorządowych prezydenci miast wybierani będą bezpośrednio przez mieszkańców. W Sejmie trwają prace nad projektem ustawy o bezpośrednim wyborze wójtów, burmistrzów i prezydentów. Ustawa ma poparcie SLD i PO. Jako kandydatów SLD na stanowisko w stołecznym ratuszu wymienia się Ryszarda Kalisza i Danutę Waniek. "Nie wyklucza" kandydowania w Warszawie Lech Kaczyński z Prawa i Sprawiedliwości. Trzecim liczącym się kandydatem może być Andrzej Olechowski, lider Platformy Obywatelskiej. Nie stanie w szranki natomiast były prezydent Paweł Piskorski.

   mT5-small: Dla nich gotowi są nawet zrezygnować z mandatów poselskich. W najbliższych tygodniach rozstrzygnie się, czy w jesiennych wyborach samorządowych prezydenci miast wybierani będą bezpośrednio przez mieszkańców. ako kandydatów SLD na stanowisko w stołecznym ratuszu wymienia się Ryszarda Kalisza i Danutę Waniek.W partii krąży opinia, że Kalisz sam się zgłosił.

   polT-small: Politycy szykują się do objęcia stanowisk prezydentów dużych miast. Dla nich gotowi są nawet zrezygnować z mandatów poselskich. W najbliższych tygodniach rozstrzygnie się, czy w jesiennych wyborach samorządowych prezydenci miast wybierani będą bezpośrednio przez mieszkańców.

   papuGaPT2: Partia ma poważne problemy z przywództwem. Nie ma mocnego nazwiska. Na razie próbuje się rozwiązywać kolejne oświęcimskie kryzysy w drodze wykorzystywania kruczków prawnych. Niedawno wyszły na jaw kontakty radnego z półświatkiem.W Szczecinie najpoważniejszym z kandydatów jest Edmund Runowicz, znany wszystkim z Gliwic lider ZChN. Z racji wpływów jakie ma partyjniactwo w mieście, chcą go popierać.Próba stworzenia drugiego centrum politycznego nie powiodła się. SLD zrezygnował z pomysłu utworzenia szerokiej koalicji ugrupowań posierpniowych. Przy budżecie państwa 181,6 mld działalność biura wyborczego kosztuje 0,041 proc. państwowych wydatków.Tymczasem w ramach walki z bezrobociem szuka ratunku w Urzędzie Miasta.

   mT5-base: W najbliższych tygodniach rozstrzygnie się, czy w jesiennych wyborach samorządowych prezydenci miast wybierani będą bezpośrednio przez mieszkańców. Tymczasem konkurenci sięgają po najlepsze kadry. Sojusz ma kłopot, bo nie ma drugiego tak dobrego kandydata. Jak nie Borowski, to kto? Jako kandydatów SLD na stanowisko w stołecznym ratuszu wymienia się Ryszarda Kalisza i Danutę Waniek. Te dwie osoby mają w stolicy duże poparcie.

   plBART: Politycy szykują się do objęcia stanowisk prezydentów dużych miast. Dla nich gotowi są nawet zrezygnować z mandatów poselskich.W najbliższych tygodniach rozstrzygnie się, czy w jesiennych wyborach samorządowych prezydenci miast wybierani będą bezpośrednio przez mieszkańców. Jako kandydatów SLD na stanowisko w stołecznym ratuszu wymienia się Ryszarda Kalisza i Danutę Waniek. "Nie wyklucza" kandydowania w Warszawie Lech Kaczyński z Prawa i Sprawiedliwości. Trzecim liczącym się kandydatem może być Andrzej Olechowski, lider Platformy Obywatelskiej.

   polT-base: Czy w jesiennych wyborach samorządowych prezydenci miast wybierani będą bezpośrednio przez mieszkańców. W Sejmie trwają prace nad projektem ustawy o bezpośrednim wyborze wójtów, burmistrzów i prezydentów. Ustawa ma poparcie głównych sił politycznych, jak SLD i PO, jest więc wielce prawdopodobne, iż zostanie uchwalona. Nic dziwnego, że już dzisiaj partie zastanawiają się, kim obsadzić najwyższe stanowiska w miastach. Dla niektórych polityków odejście do samorządu oznaczałoby konieczność złożenia mandatu posła, ale nie zraża ich to.

2. **Gold Summary:** Ludzie całkowicie sparaliżowani będą mogli samą myślą sterować komputerem, by porozumieć się z otoczeniem. Niezwykle czułe detektory, odczytujące wybrane fale mózgowe, testowane są w laboratoriach USA, Japonii i Europy. Do łączności z urządzeniami elektronicznymi

można wykorzystać jakąkolwiek aktywność układu nerwowego - sygnały elektryczne wytwarzane ruchami gałek ocznych lub impulsy nerwów mięśni.

mT5-small: Uniwersytetu Emory w Atlancie wypróbowali u dwóch pacjentów system pozwalający sterować kursorem na ekranie monitora. Sterowanie komputerem samą myślą stało się możliwe, gdy uczeni lepiej opanowali odczytywanie i werbalizację sygnałów. Komunikowanie się przy użyciu takiego systemu jest dość żmudne, ale wyłącznie dla ludzi zdrowych. Dla osób całkowicie sparaliżowanych, nie mogących wykonywać żadnych ruchów, nawet mięśniami twarzy, jest jedynym sposobem porozumiewania.

polT-small: Ludzie całkowicie sparaliżowani będą mogli samą myślą sterować komputerem, by porozumieć się z otoczeniem. Niezwykle czułe detektory, odczytujące wybrane fale mózgowe, testowane są w laboratoriach USA, Japonii i Europy. Do łączności z urządzeniami elektronicznymi można wykorzystać jakąkolwiek aktywność układu nerwowego - sygnały elektryczne wytwarzane ruchami gałek ocznych lub impulsy nerwów mięśni. Jeszcze niedawno wszelkie rozważania na ten temat były domeną wyłącznie autorów powieści fantastycznonaukowych. Obecnie naukowcy w kilkunastu ośrodkach prześcigają się w konstruowaniu urządzeń coraz sprawniej wykrywających komunikaty myślowe.

papuGaPT2: dzisiaj coraz więcej teoretyków na temat tego, co może się stać, jak będzie możliwe, za kilka lat. Trudno wyobrazić sobie wielką liczbę skomplikowaną, kręcących się po krzywej Kenii i Epsilon. Gdyby jednak wziąć pod uwagę normalny, a nie ten podkolorowany obraz powyborczej rzeczywistości, to rzeczywiście wyglądałby nędzny. Jeszcze niedawno takie pytanie byłoby niedorzeczne. Dziś naukowcy nie są już takimi sceptykami. \nNiektórzy jeszcze przed przystąpieniem Polski do UE zastanawiali się, czy w ogóle wchodzić w ten projekt. Nie wszyscy są entuzjastami nowego tworu, który stał się również przedmiotem marzeń i zazdrości milionów. A Polska? Jak wynika z sondaży, 77 proc. Amerykanów gotowych jest poprzeć likwidację lub gruntowną zmianę europejskich norm. Przeciwnicy wejścia do Unii Europejskiej obawiają się takiego myślenia, bo będzie on trwał tylko trzy i pół roku. Natomiast w rzeczywistości jest to proces bardzo długo.

mT5-base: Jeszcze niedawno wszelkie rozważania na ten temat były domeną wyłącznie autorów powieści fantastycznonaukowych. Obecnie uczeni w kilkunastu ośrodkach prześcigają się w konstruowaniu urządzeń coraz sprawniej wykrywających komunikaty myślowe. Niezwykle czułe detektory, odczytujące wybrane fale mózgowe, testowane są w laboratoriach USA, Japonii i Europy. Kilka takich rozwiązań jest testowanych w USA.

plBART: Ludzie całkowicie sparaliżowani będą mogli samą myślą sterować komputerem, by porozumieć się z otoczeniem. Niezwykle czułe detektory, odczytujące wybrane fale mózgowe, testowane są w laboratoriach USA, Japonii i Europy. Do łączności z urządzeniami elektronicznymi można wykorzystać jakąkolwiek aktywność układu nerwowego. Urządzenie ma jednak tę wadę, że wymaga użycia wszczepów, grożących powstaniem infekcji i uszkodzeniem mózgu.

polT-base: ZBIGNIEW WOJTASIŃSKI Ludzie całkowicie sparaliżowani będą mogli samą myślą sterować komputerem, by porozumieć się z otoczeniem. Niezwykle czułe detektory, odczytujące wybrane fale mózgowe, testowane są w laboratoriach USA, Japonii i Europy. Do łączności z urządzeniami elektronicznymi można wykorzystać jakąkolwiek aktywność układu nerwowego - sygnały elektryczne wytwarzane ruchami gałek ocznych lub impulsy nerwów mięśni.

I Example predictions on PSC (abstract)

1. **Gold Summary:** Anachroniczna jest koncepcja, u nas bynajmniej nierzadko wyznawana, że swoboda umów wyklucza ideę ochrony konsumenta.Przed kilkoma laty niemiecki Trybunał Konstytucyjny uznał, że nawet osoba pełnoletnia, samodzielna, nie poddana żadnemu przymusowi wymaga ochrony ręcząc za kredyt bankowy oraz że konieczna jest szczegółowa, wyczerpująca, niedwuznaczna, jasna informacja, ze wskazaniem na kwestie najbardziej niebezpieczne dla poręczyciela. Lepsza informacja dla konsumenta łagodzi bowiem nierówność pozycji rynkowej. W konsekwencji tych orzeczeń zmieniła się praktyka powszechnych sądów w Niemczech. Trzy kwestie zasługują tu na uwagę. Po pierwsze - umowy kredytowe i sytuacja poręczyciela doczekały się w Niemczech oceny TK, dokonywanej z konstytucyjnego punktu widzenia. Po drugie - TK za remedium na strukturalne zachwianie równowagi umownej uznał zwiększenie obowiązków informacyjnych kontrahenta konsumenta. Po trzecie wreszcie - opisywana sytuacja jest kolejnym przykładem tego, jak dalece anachroniczna jest koncepcja (u nas bynajmniej nierzadko wyznawana), że swoboda umów wyklucza ideę ochrony konsumenta.Z konstytucji nie można wyczytać, jakimi środkami i na jakim poziomie ma się chronić konsumenta, ale to nie jest jedyny możliwy sposób "użycia konstytucji" do takich celów. Ustawa zasadnicza da się użyć jako norma rozstrzygająca na wypadek kilku możliwych interpretacji jakiegoś przepisu: "prokonsumenckiej" (w zakresie wymienionych w konstytucji, szczególnie chronionych praw konsumenta) i "antykonsumenckiej" lub choćby "konsumencko neutralnej". Albo na wypadek interpretacji norm blankietowych, klauzul generalnych czy zwrotów niedookreślonych, które sąd musi odkodować, nadać im konkretną treść.Drugą ważną kwestią jest znaczenie informacji jako oręża konsumenta. W europejskim prawie wspólnotowym kamieniem węgielnym ochrony konsumenta jest informacja. Uważa się, że konsument wymaga ochrony, ponieważ jest źle poinformowany i na skutek tego nie może w prawdziwie wolny i nieskrępowany sposób decydować o swym "udziale na rynku". Stąd się biorą niezwykle rozbudowane w dyrektywach i ich implementacjach wewnątrzkrajowych przepisy mówiące, o czym, kiedy i jak trzeba konsumenta informować. Nie tylko jednak sama obrona przez informację jest cechą charakterystyczną europejskiego prawa konsumenckiego. Występuje tu jeszcze wskazanie minimalnego poziomu treści umowy. Oznacza to, że poszczególne państwa, implementując dyrektywy konsumenckie we własnym porządku prawnym, nie mogą zejść niżej wskazanego tam poziomu ochrony. Ale tutaj tkwi pułapka: przyznanie w prawie krajowym zbyt wysokiego poziomu ochrony, może być uznane za rzeczywistą czy ukrytą praktykę dyskryminacyjną, ograniczającą dostęp towarów i usług na rynek wspólny. Nie tylko informacja i nie tylko określenie minimalnego poziomu uprawnień zagwarantowanych konsumentowi w dotyczącym danej umowy prawie wspólnotowym czy wewnętrznym są cechami charakterystycznymi europejskiego prawa konsumenckiego. Otóż cechą charakterystyczną współczesnego prawa umów jest to, że konsumentowi przysługuje prawo "wycofania się z transakcji". W praktyce i u nas ta nowa technika ochronna zaczyna się pojawiać: zna ją polskie prawo ubezpieczeniowe. Co charakterystyczne: tego rodzaju możliwość wycofania się z transakcji służy (w prawie europejskim) jedynie konsumentowi, nie jego kontrahentowi. I znów jest to jeden ze środków, który przyznaje się tylko stronie instytucjonalnie słabszej, używany, by zrównoważyć siłę rynkową konsumenta wobec profesjonalisty.

   mT5-small: Pod koniec laty dwa orzeczenia niemieckiego Trybunału Konstytucyjnego zbulwersowały światek prawniczy. Trybunał uznał, że anachroniczna jest koncepcja, u nas bynajmniej nierzadko wyznawana, jako swoboda umów wyklucza ideę ochrony konsumenta. Po trzecie - umowy kredytowe i sytuacja poręczyciela doczekały się w Niemczech oceny TK, dokonywanej z konstytucyjnym punktu widzenia.

   polT-small: Przed kilkoma laty dwa orzeczenia niemieckiego Trybunału Konstytucyjnego zbulwersowały świat prawniczy. Trybunał uznał, że nawet osoba pełnoletnia, samodzielna i niepoddana żadnemu przymusowi wymaga ochrony ręcząc za kredyt bankowy. Drugą istotną kwestią europejskiego prawa konsumenckiego jest wskazanie minimalnej wiedzy i doświadczenia między poręczycielem i bankiem. Innym przykładem środka ochrony konsumenta jest niemiecki Trybunał

Konstytucyjny. Moim zdaniem, konsumentowi przysługuje prawo wycofania się z transakcji bez negatywnych konsekwencji.

papuGaPT2: W porównaniu z prawem polskim ustawodawstwo i orzecznictwo krajów europejskich znacznie bardziej chroni prawa konsumentów, co w konsekwencji doprowadzić może do unikania przez banki i inne instytucje finansowe prokonsumenckie wszelkiej informacji o towarze i bezpiecznym sposobie korzystania z niego. Europejskie dyrektywy konsumenckie wskazują minimalny poziom treści umowy: są to przepisy zakazujące reklamy pewnych produktów czy usług (lekarstwa, alkohole, wolne zawody), używania zrozumiałego w określonych mediach (radio, tv), reklamy kierowanej do pewnych osób (dzieci) czy też zniósł dobrze ochronę przed wprowadzającymi w błąd informacjami na etykietach. Wreszcie cechą charakterystyczną współczesnego prawa umów jest to, iż konsument przez krótki czas (siedem do dziesięciu dni) ma prawo, bez żadnych negatywnych konsekwencji, zrezygnować z transakcji. Środek ten przysługuje tylko konsumentowi jako stronie instytucjonalnie słabszej; jego kontrahenta obowiązuje ogólna zasada mówiąca o "umowy powinny zostać dotrzymywane". Zasada ta wyraża się w łacińskiej formule "in dubio contra proferentem" i jest ważnym instrumentem prawa cywilnego, który służy nie tylko ochronie konsumenta, ale także jako środek prawnokarnego (art. 76), który ma zrównoważyć siłę przyciągania i działać na rzecz zrekompensowania braków jego wiedzy. Stąd się biorą rozbudowane przepisy mówiące, o czym, kiedy i jak należy informować konsumenta o wszelkich aspektach umowy kredytowej, a także o możliwych niebezpieczeństwach i ryzykach, które mogą być nieświadomym wyrazem orientacji konsumenta w transakcjach rynkowych.

mT5-base: Rę swoboda umów wyklucza ideę ochrony konsumenta.

plBART: Przed kilkoma laty dwa orzeczenia niemieckiego Trybunału Konstytucyjnego zbulwersowały świat prawniczy. Trybunał uznał, że nawet osoba pełnoletnia, samodzielna, nie poddana żadnemu przymusowi wymaga ochrony ręcząc za kredyt bankowy, bo nierównowaga wiedzy i doświadczenia między poręczycielem i bankiem jest zbyt wielka i tolerowanie jej zagraża jednej z konstytucyjnych zasad niemieckiego porządku prawnego: wyrażonej w art. 2 konstytucji autonomii woli. W konsekwencji orzeczeń TK zmieniła się praktyka niemieckich sądów, które umowy kredytowe doczekały się oceny TK, dokonywanej z konstytucyjnego punktu widzenia. Po drugie, TK uznał za remedium na strukturalne zachwianie równowagi umownej. Opisywana sytuacja jest kolejnym przykładem tego, jak dalece anachroniczna jest koncepcja (u nas bynajmniej nierzadko wyznawana), że swoboda umów wyklucza ideę ochrony konsumenta.

polT-base: W.

2. **Gold Summary:** Praca Huberta Orłowskiego "Polnische Wirtschaft. Nowoczesny niemiecki dyskurs o Polsce" poświęcona jest stereotypom dotyczącym stosunków polsko-niemieckich, w tym jednemu z najsilniejszych - "polnische Wirtschaft". Autorstwo tego stereotypu przypisuje się Georgowi Fosterowi, uczonemu i podróżnikowi, który w swoich listach pisanych pod koniec XVIII w. terminem tym określał kombinację bałaganu, nieudolności, brudu i lenistwa. W stereotyp myślenia Niemców o Polsce wpisywały się wydarzenia historyczne. Zrywy niepodległościowe Polaków od powstania listopadowego do "Solidarności" traktowane były często jako przejaw polskiej skłonności do anarchii. We współczesnych Niemczech stereotyp "polnische Wirtschaft" jest wciąż żywy. Podtrzymują go zarówno media, jak i w literatura piękna.

   mT5-small: Nakładem wydawnictwa Huberta Orłowskiego poświęcona jest stereotypom dotyczącym stosunków polsko-niemieckich, ze szczególnym uwzględnieniem najsilniejszych z nich: „polnische Wirtschaft”. Autorstwo tego stereotypu przypisuje się niejakiemu Georgowi Fosterowi, uczonemu i podróżnikowi z Żuław Wiślanych. "Polnische Wirtschaft" oznaczającemu kombinację tych cech. Wśród warstw bogatszych widziano rozrzutność, równocześnie jednak także brak higieny, zacofanie, ignorancja.

   polT-small: W książce profesora Huberta Orłowskiego poświęcona jest stereotypom dotyczącym stosunków polsko-niemieckich, ze szczególnym uwzględnieniem najsilniejszych z nich: "polnische

Wirtschaft" oznaczającemu kombinację bałaganu, nieudolności, brudu i złą gospodarkę.

papuGaPT2: Od lat 40. literatura podtrzymywała mit szlachetnego Polaka, o którym zapomina młodsze pokolenie. Stereotyp ten nigdy by może nie zaistniał, gdyby nie głębokie różnice dzielące polskie cnoty szlacheckie od niemieckich cnót mieszczańskich, pogłębione różnicami między katolicyzmem i protestantyzmem. Po klęsce powstania listopadowego, po upadku muru berlińskiego okazało się, że ustrój ten nie był w stanie zdobyć uznania i poparcia ludu, dlatego też w książkach, operetkach i filmach pełno jest dowcipów o polskiej gospodarce i polskim sejmikowaniu. Profesor Hubert Orłowski jest współautorem tzw. ustawy teczkowej, ustanawiającej Instytut Pamięci Narodowej, oraz ustawy o ochronie informacji niejawnych, określającej m.in. mianem "pacta servanda sunt" na której spoczywa cała odpowiedzialność.

mT5-base: W niemieckich zbiorów przysłów o Polsce, ukazała się dwa lata temu w Wiesbaden nakładem wydawnictwa Otto Harrasowitza na polski.

plBART: Praca profesora Huberta Orłowskiego "Polnische Wirtschaft. Nowoczesny niemiecki dyskurs o Polsce" poświęcona jest stereotypom dotyczącym stosunków polsko-niemieckich, ze szczególnym uwzględnieniem najsilniejszych z nich: „polnischer Wirthschaft”. Termin ten określa "nieporządek i złą gospodarkę, niedbalstwo i bałagan". Autorstwo tego, mającego długą historię i niesłychaną żywotność, stereotypu przypisuje się niejakiemu Georgowi Fosterowi, uczonemu i podróżnikowi z Żuław Wiślanych. Naprawdę jednak formuła była już gotowa i nieźle znana wcześniej, choć w myśleniu Niemców (i nie tylko ich) zadomowiła się dopiero ok. 1830 r. po edycji listów Fostera. W liście do swojego wierzyciela pisał z Wilna w 1784 roku: "O nieopisanym brudzie, lenistwie, opilstwie i nieudolności całej służby..., o niezdarności rzemieślników, ich niesłychanie kiepskiej robocie, wreszcie o zadowoleniu Polaków [Polaken] z własnego bagienka, a także ich przywiązaniu do rodzinnych zwyczajów nie chcę pisać już nic więcej, aby nie przedłużać tego listu".

polT-base: W.