Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4374–4394
Marseille, 20-25 June 2022

Evaluation of Transfer Learning for Polish with a Text-to-Text Model

Aleksandra Chrabrowa, Łukasz Dragan, Karol Grzegorczyk, Dariusz Kajtoch
Mikołaj Koszowski, Robert Mroczkowski, Piotr Rybak
Allegro ML Research at Allegro.pl,
ul. Grunwaldzka 182, 60-166 Poznań, Poland
{firstname.lastname}@allegro.pl

Abstract

We introduce a new benchmark for assessing the quality of text-to-text models for Polish. The benchmark consists of diverse tasks and datasets: KLEJ benchmark adapted for text-to-text, en-pl translation, summarization, and question answering. In particular, since summarization and question answering lack benchmark datasets for the Polish language, we describe their construction and make them publicly available. Additionally, we present plT5 - a general-purpose text-to-text model for Polish that can be fine-tuned on various Natural Language Processing (NLP) tasks with a single training objective. Unsupervised denoising pre-training is performed efficiently by initializing the model weights with a multi-lingual T5 (mT5) counterpart. We evaluate the performance of plT5, mT5, Polish BART (plBART), and Polish GPT-2 (papuGaPT2). The plT5 scores top on all of these tasks except summarization, where plBART is best. In general (except for summarization), the larger the model, the better the results. The encoder-decoder architectures prove to be better than the decoder-only equivalent.

Keywords: benchmark, corpus, evaluation, NLU, NLG, T5, Polish language, summarization,

1. Introduction

Recent years have brought significant progress in 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 problems 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 the same architecture, training procedure, and decoding process for many NLP tasks such as classification, machine translation, summarization, and question answering, to name a few. In addition, Raffel et al. (2020) demonstrated that the simplicity of this approach combined with scale could achieve state-of-the-art results on many benchmark datasets, which makes it even more attractive. 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. Multiple publications show 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:

- 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,
- construction of benchmark datasets in the Polish domain for question answering and summarization,
- demonstration of the efficiency of pre-training procedure for transferring knowledge from multilingual to monolingual text-to-text models based on work by (Arkhipov et al., 2019; Mroczkowski et al., 2021),
- release of plT5[3] – a T5-based model for the Polish language, which achieves the best results among the evaluated text-to-text models on KLEJ benchmark, machine translation and question answering and second-best results in summarization.

2. Polish T5 model

There was no large-scale text-to-text model for Polish to benchmark transfer learning methods, so we trained

[1] https://github.com/sdadas/polish-nlp-resources#bart
[2] https://hf.co/flax-community/papuGaPT2
[3] https://hf.co/allegro/plt5-base

three versions of plT5 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). We addressed the difference in tokenizers' vocabulary as described in (Arkhipov et al., 2019; Mroczkowski et al., 2021). We copied embedding weights for tokens appearing in both dictionaries directly. Otherwise, if the target vocabulary token was missing in the source vocabulary, it was split into subtokens, and the embedding was the average of subtokens embeddings.

2.3. Pre-training objective

For the training of the plT5 model, we leveraged the self-supervised denoising objective proposed by (Raffel et al., 2020) i.e. span-corruption. The model was fed with the corrupted version of the original sentence during training. Single sentinel tokens replaced the consecutive spans of input tokens, and the target was to reconstruct replaced tokens.

2.4. Pre-training datasets

Denoising pre-training was performed on a weighted mixture[4] of the following corpora: Polish Wikipedia, National Corpus of Polish (NKJP, Przepiórkowski et al. (2011)), Wolne Lektury[5], Polish Open Subtitles[6] and Common Crawl of Polish sites. Each dataset is described in (Mroczkowski et al., 2021).

2.5. Tokenizer

The training dataset was tokenized into subword units using a sentencepiece unigram model (Kudo, 2018) with a vocabulary size of 50k tokens. Unigram model was trained[7] on the most representative parts of our corpus, i.e., the subset (random 25% of the whole corpus) of the NKJP, and Polish Wikipedia.

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

[5]https://wolnelektury.pl/

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

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

2.6. Pre-training procedure

For self-supervised denoising pre-training, we used the original T5 training scripts[8], which are based on the Mesh TensorFlow framework (Shazeer et al., 2018). In addition, we used default options for the span-corruption objective with a mean span length of 3 and a corruption rate of 15% (Raffel et al., 2020).

plT5 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 a simple hypertuning procedure. We evaluated plT5-base model snapshots saved after 10k training steps for pre-training learning rates in {1e-3, 2e-3, 5e-3}. We used KLEJ benchmark tasks equipped with validation sets for performance evaluation: AR, NKJP-NER, PolEmo2.0-OUT, CDSC-E. Best model training was continued up to 50k steps.

3. Downstream tasks

plT5 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 a text-to-text reformulation of the 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 and papuGaPT2 to compare their performance. The trainable number of parameters and vocabulary size for each architecture is shown in Table 1.

Model		# Params	Vocab. size
Small models	mT5	300M	250k
	plT5	95M	50k
Base models	mT5	582M	250k
	plT5	275M	50k
	plBART	139M	50k
	papuGaPT2	124M	50k
Large models	mT5	1.23B	250k
	plT5	820M	50k

Table 1: Summary of the 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 the KLEJ benchmark (Rybak et al., 2020). We used all of the given tasks to compare the generative models, except

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

for the CDSC-R regression. We formed the tasks into a text-to-text format proposed in (Raffel et al., 2020). During the fine-tuning phase, the model was provided descriptive input containing the specific prefix and the textual features with descriptive labels[9]. 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. Task definition

We transformed KLEJ tasks into a text-to-text format where each input sentence (one or two in KLEJ tasks) has 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 a 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 in Table 2.

3.1.2. Task 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.

3.1.3. Experiments

We followed a simple experimental setup to ensure a fair comparison between models. The learning rate was the only hyperparameter we tuned. 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. 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 ones.

3.1.4. Results

The results for KLEJ benchmark (text-to-text format) are in Table 3. In general, encoder-decoder models dedicated to the Polish language perform significantly better at tasks in the KLEJ benchmark than their multilingual counterparts. The most striking performance difference occurs for the base version of the model on the PolEmo2.0-OUT task. The difference between plT5-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 multilingual 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 plT5 model in each task. The best encoder-decoder model is plT5-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, HerBERT-large. The difference between these architectures is, on average, 1pp. We observe the most significant performance gap of 6.8pp when moving from plT5-small to plT5-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 plT5-base. It is worth emphasizing that almost two 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 par with plT5-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 the 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 original text are not forbidden).

3.2.1. Datasets

Allegro Articles (AA) Collection of over 33k articles from Allegro.pl[10] - 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[11]: (1) *body2lead* - generate lead from a body of the article (2) *body+lead2title* -

[9] We did not explore multi-task learning in this setup with task-specific prompts

[10] Oryginal articles are available at https://allegro.pl/artykuly

[11] Prepared datasets are available at https://hf.co/datasets/allegro/summarization-allegro-articles

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 2: Prompts and labels used in the 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.

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
	plT5	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
	plT5	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
	plT5	**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 3: Evaluation results on KLEJ benchmark. *AVG* is the average score across all tasks. Scores are reported for the test set and correspond to median values across seven runs for small and base models and three runs for large versions 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. F1 scores are reported for CBD, Czy wiesz? and PSC, Spearman correlations are reported for CDSC-R, MAE based score (Rybak et al., 2020) for AR, and accuracy scores are reported for the other tasks.

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[12]: (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.

3.2.2. Experiments

To compare the models fairly, we designed an experimental setup where the learning rate is the only hyperparameter (details in Appendix B.4). By default, each model was trained with three seeds for ten 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 (PSC task) trained for four epochs.

Trimming inputs Summarization models by its nature deal with long source texts and shorter target texts.

[12]Prepared datasets are available at https://hf.co/datasets/allegro/summarization-polish-summaries-corpus

Unfortunately, we had to trim substantially 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) were affected the most. The models never saw 100% of source text during training, but between 26-52%. Details are in Appendix B.3.

3.2.3. Results

Results are shown in Table 4 which contains arithmetic mean of (f-measure) ROUGE-1, ROUGE-2 and ROUGE-L (Lin, 2004)[13] for each model and task. More detailed metrics for each task can be found in Appendix C (Tables 11-15). In all cases, the best performing model was plBART, with one exception of AA *body+lead2title* task in which plT5-large outperformed the others. In terms of the average performance over all the tasks, plBART is the winner, and the second is plT5-small. plT5 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 2 baselines: copying $n = 3$ leading sentences from source as candidate target summarization, following (Gliwa et al., 2019; See et al., 2017) and much better *adaptive n* baseline: copying n leading sentences from source that best match the average target length. 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).

Upper bound estimates PSC dataset contains multiple summaries of different lengths and by different authors per source text. We evaluate human performance as ROUGE metrics between different summaries for the same source text. 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, particularly for abstractive summaries. This is in agreement with the systematic evaluation of summarization metrics by (Fabbri et al., 2021).

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. If the source was trimmed down significantly, 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.

After exploration, we found out that, on average, summaries generated by encoder-decoder architectures (plT5, mT5, plBART) are shorter than targets. 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. Furthermore, the PSC abstract was particularly difficult for plT5-base which generated extremely short summaries (samples are in Appendix F, G, H).

3.3. Question answering

Question answering (QA) systems enable users to automatically obtain accurate answers to natural language questions. 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 the question answering system for the Polish language, so we 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 noticed many low-quality questions that cannot be answered using Wikipedia as a knowledge base. Therefore, we removed questions from specialized domains (e.g., TV series, movies) with open-ended answers ("why?" and "how?" questions), lacking the answer or the answer depending on when the question is asked (e.g. "Who won World Cup this year?"). After filtration, we left 1875 valuable questions.

Jeden z Dziesieciu 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[14].

Poleval is an evaluation campaign for NLP tools for Polish. The 2021 edition contains the question answer-

[13] 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.

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

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
	plT5	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**
	plT5	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
	plT5	18.9 ± 1.8	12.1	**39.4**	17.5	15.1	10.6

Table 4: Evaluation results on all summarization tasks (test split, mean values across three runs). For simplicity only arithmetic mean ROUGE (f-measure) is reported, ROUGE-1, ROUGE-2 and ROUGE-L are in Appendix C (Tables 11-15). The best scores are in bold, and results above the baseline are underlined. ROUGE AVG is the arithmetic mean of mean ROUGE of all tasks per model.

ing task[15] 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 the test. The test-B set was not released at the time of conducting the experiments.

Final dataset combines the datasets mentioned above, 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 pretrained models. Therefore, 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 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 was 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 trained 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[16]. We combined them with "Czy wiesz?" dataset (Marcinczuk et al., 2013; Rybak et al., 2020) to train a neural retriever based on the 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 fine-tuned the following models: plT5-base, plT5-wiki[17], plT5-large, mT5-base, plBART, papuGaPT2, and T5 model initialized 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 plT5-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.

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

[16] We release the dataset at https://hf.co/datasets/allegro/polish-question-passage-pairs

[17] plT5-base additionally fine-tuned using Wikipedia corpus

We used a single NVIDIA A100 GPU for all the models. The best results for each model are presented in Table 5. In both tasks, the best model was plT5-large. In base size, plT5-wiki performs the best in both tasks. Models trained on open-ended tasks 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 a similar domain (e.g., it answers "Montmartre" for the question of "the highest peak of the Beskids range"). Fine-tuning 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 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 the summarization task 3.2.2. Because of papuGaPT2 input size limitation, we could not pass the entire passage to the model. Therefore, we repeated plT5-wiki and plBART training using a 1024 input size setting. The results for plT5 and plBART were not significantly different from those with no size-limited input, which confirms their superiority over decoder-only architecture in this task.

Task		open	passages
Base models	plT5	20.9	42.8
	plT5-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	plT5	**26.5**	**51.3**

Table 5: 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 plT5 on parallel corpora consisting of pairs of English and Polish sentences and evaluated 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[18] (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).

3.4.2. Experiments

We evaluated the models on the WMT20 development set using the BLEU score (Papineni et al., 2002). For the generation, we used beam search with five beams and a maximal sequence length of 100. Input sentences were also limited to 100 tokens. We compared plT5, mT5, plBART, and papuGaPT2. The models were trained in both directions simultaneously, with examples fed to the models alternately. Since the plT5 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 plT5 tokenizer, we applied a technique based on Arkhipov et al. (2019, section 3). We applied the same conversion to those models to have a fair comparison with mT5, plBART, and papuGaPT2. 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 the 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 Table 6. We used a single NVIDIA A100 GPU with 40 GB of memory for all the models. Small and base

Model		Vocab	Batch size
Small models	mT5	mT5	25
	plT5	plT5	50
	mT5	wmt20	50
	plT5		
Base models	mT5	mT5	10
	plT5	plT5	25
	mT5	wmt20	25
	plT5		
	plBART	plBART wmt20	40
	papu-GaPT2	papuGaPT2 wmt20	25
Large models	mT5	wmt20	10
	plT5		

Table 6: Machine translation batch sizes

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

models were trained for five epochs and large for two.

3.4.3. Results

The results are presented in Table 7. plT5 generates translations superior to mT5 in direction en→pl but not in pl→en. This may be because plT5 was trained on Polish-only data and, therefore, the ability to generate English text deteriorates. plT5 also outperforms plBART and papuGaPT2. We repeated plT5-base and plBART experiments for three different seeds. The variance of the plT5 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 plT5 are not up to the level of the WMT20 winning team's results (Krubiński et al., 2020). However, plT5-large performance in the en→pl direction is slightly better than their bi-directional baseline. However, in this work, we did not attempt to achieve state-of-the-art on those datasets but to compare plT5 with similar architectures in similar settings.

4. Conclusion

In this work, we introduced a benchmark for natural language generation in Polish composed of 4 tasks constructed from publicly available datasets, namely the KLEJ benchmark adapted for text-to-text, en-pl translation, summarization with varying levels of abstractedness, and question answering. In addition, we introduced a novel text-to-text model for Polish, plT5. 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 attribute this performance to the pre-training scheme where BART reconstructs the entire input in the decoder and thus possesses *copy bias*. We should state that the overall performance of plBART is impressive, considering that it has almost twice as few parameters as plT5-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 pre-trained plT5 by initializing the weights from mT5 checkpoints with no exhaustive training. We made our models and new generation benchmark publicly available.

We did not explore the possibility of optimizing prefixes and labels for the text-to-text KLEJ benchmark and leave for future work. Thus, results reported in Table. 3 need to be treated as the lower bound. Additionally, our summarization dataset includes summaries of different lengths. This introduces additional ambiguity, especially during evaluation, since the model does not know what should be the target length of the summary. It would be interesting to check if conditioning the text-to-text model on summary length would lead to better results.

Model		Vocab	en→pl	pl→en
Small models	mT5	mT5	20.5	25.0
	plT5	plT5	20.3	24.7
Base models	mT5	wmt20	20.8	24.8
	plT5		20.8	25.4
	mT5	mT5	22.4	27.0
	plT5	plT5	23.2	26.7
	mT5	wmt20	22.5	26.7
	plT5		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	**29.0**
	plT5		**25.5**	28.9

Table 7: BLEU scores for the WMT20 devset. The best results in each category (small, base, and large) are underlined. The best overall results are shown in bold.

Acknowledgments

We thank Alina Wróblewska from Institute of Computer Science, Polish Academy of Sciences for help with preprocessing the National Corpus of Polish.

References

Arkhipov, M., Trofimova, M., Kuratov, Y., and Sorokin, A. (2019). Tuning multilingual transformers for language-specific named entity recognition. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 89–93, Florence, Italy, August. Association for Computational Linguistics.

Barrault, L., Biesialska, M., Bojar, O., Costa-jussà, M. R., Federmann, C., Graham, Y., Grundkiewicz, R., Haddow, B., Huck, M., Joanis, E., Kocmi, T., Koehn, P., Lo, C.-k., Ljubešić, N., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Pal, S., Post, M., and Zampieri, M. (2020). Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online, November. Association for Computational Linguistics.

Carmo, D., Piau, M., Campiotti, I., Nogueira, R., and Lotufo, R. (2020). Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. *arXiv preprint arXiv:2008.09144*.

Caswell, I., Kreutzer, J., Wang, L., Wahab, A., van Esch, D., Ulzii-Orshikh, N., Tapo, A., Subramani, N., Sokolov, A., Sikasote, C., et al. (2021). Quality at a glance: An audit of web-crawled multilingual datasets. *arXiv preprint arXiv:2103.12028*.

Chan, B., Schweter, S., and Möller, T. (2020). German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (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, June. Association for Computational Linguistics.

Fabbri, A. R., Kryściński, W., McCann, B., Xiong, C., Socher, R., and Radev, D. (2021). SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 04.

Gliwa, B., Mochol, I., Biesek, M., and Wawer, A. (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, November. Association for Computational Linguistics.

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

Jónsson, H. P., Símonarson, H. B., Snæbjarnarson, V., Steingrímsson, S., and Loftsson, H. (2020). Experimenting with different machine translation models in medium-resource settings. In Petr Sojka, et al., editors, *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.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koupaee, M. and Wang, W. Y. (2018). Wikihow: A large scale text summarization dataset.

Krubiński, M., Chochowski, M., Boczek, B., Koszowski, M., Dobrowolski, A., Szymański, M., and Przybysz, P. (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, November. Association for Computational Linguistics.

Kudo, T. (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, July. Association for Computational Linguistics.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., and Petrov, S. (2019). Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, March.

Le, H., Vial, L., Frej, J., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L., and Schwab, D. (2020). FlauBERT: Unsupervised language model pre-training for French. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2479–2490, Marseille, France, May. European Language Resources Association.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (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, July. Association for Computational Linguistics.

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

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., and Zettlemoyer, L. (2020). Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Longpre, S., Lu, Y., and Daiber, J. (2020). Mkqa: A linguistically diverse benchmark for multilingual open domain question answering.

Marcinczuk, M., Ptak, M., Radziszewski, A., and Piasecki, M. (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*.

Martin, L., Muller, B., Ortiz Suárez, P. J., Dupont, Y., Romary, L., de la Clergerie, É., Seddah, D., and Sagot, B. (2020). CamemBERT: a tasty French language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7203–7219, Online, July. Association for Computational Linguistics.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S.,

and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Mroczkowski, R., Rybak, P., Wróblewska, A., and Gawlik, I. (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, April. Association for Computational Linguistics.

Nagoudi, E. M. B., Chen, W.-R., Abdul-Mageed, M., and Cavusoglu, H. (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, June. Association for Computational Linguistics.

Nagoudi, E. M. B., Elmadany, A., and Abdul-Mageed, M. (2021b). Arat5: Text-to-text transformers for arabic language understanding and generation.

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

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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, July. Association for Computational Linguistics.

Pennington, J., Socher, R., and Manning, C. (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, October. Association for Computational Linguistics.

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

Przepiórkowski, A., Bańko, M., Górski, R. L., Lewandowska-Tomaszczyk, B., Łaziński, M., and Pezik, P. (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.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Roberts, A., Raffel, C., and Shazeer, N. (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, November. Association for Computational Linguistics.

Rybak, P., Mroczkowski, R., Tracz, J., and Gawlik, I. (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, July. Association for Computational Linguistics.

See, A., Liu, P. J., and Manning, C. D. (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, July. Association for Computational Linguistics.

Shazeer, N. and Stern, M. (2018). Adafactor: Adaptive learning rates with sublinear memory cost.

Shazeer, N., Cheng, Y., Parmar, N., Tran, D., Vaswani, A., Koanantakool, P., Hawkins, P., Lee, H., Hong, M., Young, C., et al. (2018). Mesh-tensorflow: deep learning for supercomputers. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 10435–10444.

Virtanen, A., Kanerva, J., Ilo, R., Luoma, J., Luotolahti, J., Salakoski, T., Ginter, F., and Pyysalo, S. (2019). Multilingual is not enough: Bert for finnish.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (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, June. Association for Computational Linguistics.

Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Abrego, G. H., Yuan, S., Tar, C., Sung, Y.-H., et al. (2019). Multilingual universal sentence encoder for semantic retrieval. *arXiv preprint arXiv:1907.04307*.

Zeng, C., Li, S., Li, Q., Hu, J., and Hu, J. (2020). A survey on machine reading comprehension: Tasks, evaluation metrics and benchmark datasets.

A. Summary of the released datasets

We released six new datasets. All of them are available at https://hf.co/allegro. Their sizes are shown in Table 8.

Dataset name	Examples		
	Train	Dev	Test
AA body2lead	380351	41966	104514
AA body+lead2title	380351	41966	104514
PSC whole	512201	57180	141293
PSC extract	413139	45413	112880
PSC abstract	100043	11222	27976
Polish Q-P Pairs.	10429		

Table 8: Sizes of the released datasets.

B. Summarization task

Level of abstractedness Abstractedness of the dataset is measured by calculating the unique n-grams in the reference summary, which are not in the article (Koupaee and Wang, 2018).

Compression ratio The compression ratio is defined as the ratio between the average length of the source and the average length of summaries.

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 the category tree, and other metadata. In the 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 two versions of the abstractive summarization task, which affect the summary ratio. In *body2lead* task, we use the 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 a 80% train set and 20% test set in a stratified way using a top-level category in the Allegro category tree as a label. Then, we saved 10% of the train set for validation (also stratified split) and hyperparameter tuning. Table 9 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[19]. 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, 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 9 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		
Compression ratio					
-	13.0	82.0	8.2		
Level of abstractedness					
1-grams	0.53	0.30	0.05	0.01	0.23
2-grams	0.87	0.69	0.16	0.07	0.48
3-grams	0.95	0.85	0.25	0.15	0.61
4-grams	0.97	0.92	0.32	0.22	0.68
5-grams	0.98	0.95	0.38	0.28	0.73
Number of examples					
Train	24.4k		7.8k	6.1k	1.7k
Dev	2.7k		0.9k	0.7k	0.2k
Test	6.8k		2.2k	1.7k	0.5k

Table 9: Overview of summarization tasks with the total number of examples for each split and average length (in characters) of source and target

[19] We used publicly available version https://hf.co/datasets/polsum

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 could not 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 the 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 few 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 10. 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 plT5 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.

Task		PSC	AA body2lead	AA body+lead2title
Small models	mT5	33%	85%	80%
	plT5	51%	96%	95%
Base models	papuGaPT2	26% (512)	92% (767)	96% (959)
	mT5	33%	85%	80%
	plBART	52%	96%	95%
	plT5	51%	96%	95%
Large models	mT5	25% (768)	85%	80%
	plT5	51%	96%	95%

Table 10: 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.

B.4. Learning rate tuning

On the 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 the PSC task, we performed a learning rate search for each model on the PSC whole subset and used the same learning rates for the PSC extract and PSC abstract subsets. The search was geometrical, with the start point at 4e-4, i.e., at the first subset {2e-4, 4e-4, 8e-4} was examined, each learning rate with three seed runs. The search was completed if the middle point was better than the two others. 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 a result, learning rates from 5e-5 to 2.56e-2 were selected for different models. The only exception was the 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. Summarization ROUGE scores

Model		ROUGE-1	ROUGE-2	ROUGE-L	ROUGE
Baselines lead source sentences	n=3	**20.3**	**2.8**	**14.1**	**12.4**
	adaptive	20.2	**2.8**	**14.1**	**12.4**
Small models	mT5	19.7	4.1	15.2	13.0
	plT5	21.5	4.9	16.5	14.3
Base models	papuGaPT2	20.2	1.6	12.6	12.1
	mT5	21.4	4.6	16.3	14.1
	plT5	22.3	5.2	17.1	14.9
	plBART	**24.0**	**5.1**	**17.7**	**15.6**
Large models	mT5	16.7	2.7	13.4	10.9
	plT5	19.2	2.7	14.4	12.1

Table 11: 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 source sentences	n=3	9.3	2.9	8.1	6.8
	adaptive	**10.4**	**3.6**	**9.7**	**7.9**
Small models	mT5	39.3	25.1	38.2	34.2
	plT5	41.4	27.2	40.1	36.3
Base models	papuGaPT2	18.8	5.2	18.0	14.0
	mT5	41.3	27.0	40.1	36.1
	plBART	44.0	29.6	41.3	38.3
	plBART†	44.0	29.7	42.4	38.7
	plT5	44.1	29.7	42.8	38.9
	plT5†	43.9	29.4	42.6	38.7
Large models	mT5	38.4	24.7	37.3	33.5
	plT5	**44.6**	**30.2**	**43.3**	**39.4**

Table 12: 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 source sentences	n=3	28.2	16.5	21.3	22.0
	adaptive	**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
	plT5	36.7	24.3	30.4	30.5
Base models	papuGaPT2	26.9	7.6	15.5	16.7
	mT5	27.4	14.1	20.3	20.6
	plBART	**39.0**	**26.3**	**32.4**	**32.6**
	plT5	32.9	18.1	24.5	25.2
Large models	mT5	20.6	3.7	11.5	12.0
	plT5	23.1	12.0	17.4	17.5

Table 13: 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 source sentences	n=3	29.5	18.3	22.6	23.4
	adaptive	**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
	plT5	27.7	18.3	23.6	23.2
Base models	papuGaPT2	26.9	8.5	15.9	17.1
	mT5	27.6	15.0	20.9	21.2
	plBART	**39.9**	**28.6**	**34.4**	**34.3**
	plT5	31.9	18.0	24.1	24.7
Large models	mT5	19.3	2.6	10.9	10.9
	plT5	22.4	7.3	15.6	15.1

Table 14: 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 source sentences	n=3	27.1	14.2	19.7	20.3
	adaptive	**35.4**	**17.8**	**23.6**	**25.6**
Upper bounds	human performance	34.6	16.7	25.0	25.4
Small models	mT5	28.0	14.9	20.7	21.2
	plT5	30.4	16.5	22.7	23.0
Base models	papuGaPT2	26.0	5.7	13.9	15.2
	mT5	12.0	5.0	9.5	8.8
	plBART	**33.5**	**17.7**	**25.4**	**25.5**
	plT5	15.6	7.7	12.0	11.7
Large models	mT5	16.1	1.9	9.4	9.1
	plT5	15.6	5.1	11.2	10.6

Table 15: 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.

D. Example predictions on Allegro Reviews *body+lead2title*

1. **Gold Summary:** Speedminton – badminton na wietrzne dni

   plT5-base: Jak zaczać grać w speedminton?

   plBART: Speedminton – sport dla każdego

   papuGaPT2: Squa – jak zaczać?

   mT5-base: Jak zaczać grać w speedmintona?

   plT5-large: Speedminton – sport dla aktywnych

2. **Gold Summary:** Urzadzamy wnetrze po skandynawsku

   plT5-base: Jak urzadzić mieszkanie w stylu skandynawskim?

   plBART: Jak urzadzić mieszkanie w stylu skandynawskim?

   papuGaPT2: Skandynawski styl w wersji budżetowej – jak go urzadzić?

   mT5-base: Jak urzadzić wnetrze w stylu skandynawskim?

   plT5-large: Jak urzadzić mieszkanie w stylu skandynawskim?

3. **Gold Summary:** Czy warto robić domowe peelingi – poradnik

   plT5-base: Najlepsze peelingi do domu

   plBART: Najlepsze peelingi do twarzy

   papuGaPT2: Peeslingi – najlepszy sposób na przesuszona skóre

   mT5-base: Jak zrobić peelingi w domu?

   plT5-large: Jak zrobić peeling w domu?

4. **Gold Summary:** Szerokie spodnie – must have sezonu wiosna/lato 2015

   plT5-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

   plT5-large: Wiosenne spodnie w stylu lat 70.

5. **Gold Summary:** Przygotuj sie na lodowisko

   plT5-base: Jak rozpocząć sezon na łyżwy?

   plBART: Jak zaczać przygode z łyżwami?

   papuGaPT2: Jakie ubrania na sezon kapielowy?

   mT5-base" Jak zaczać sezon na łyżwy?

   plT5-large: Jak przygotować sie do sezonu na łyżwy?

6. **Gold Summary:** Wózek widłowy niczym limuzyna sterowana iPadem

   plT5-base: Wózki widłowe zwiekszajace możliwości swojego działania

   plBART: Wózki widłowe – zwiekszamy wydajność i oszczedzamy paliwo

   papuGaPT2: Nowoczesny wózek widłowy – dlaczego jest tak ważny?

   mT5-base: Czym sa nowoczesne technologie w wózkach widłowych?

   plT5-large: Automatyzacja wózków widłowych

E. Example predictions on Allegro Reviews *body2lead*

1. **Gold Summary:** Każdy czuje respekt przed królowa nauk. Matematyka spedza sen z powiek niejednemu uczniowi. Sa jednak sposoby zachecenia dziecka do nauki matematyki i sprawienia, że lepiej ja zrozumie i polubi.

   plT5-base: Matematyka to jedna z najprzyjemniejszych dziedzin nauki. Uczymy sie ja w szkole i w domu. Nauka przez zabawe może być świetna rozrywka dla całej rodziny. Jakie gadżety moga pomóc dziecku w nauce liczenia?

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

   papuGaPT2: Nauka poprzez zabawe to jeden z najlepszych sposobów na rozwijanie zdolności poznawczych i umiejetności logicznego myślenia. Podpowiadamy, co warto zabrać ze soba do szkoły i na co zwrócić uwage przy wyborze odpowiednich akcesoriów.

   mT5-base: Nauka matematyki to doskonały sposób na spedzenie wolnego czasu w gronie najbliższych. Dzieci chetnie siegaja po pomoce naukowe, które zacheca do nauki przez zabawe i doświadczenie.

   plT5-large: Matematyka nie musi być nudna. Nauka liczenia może być wspaniała zabawa, która wspomoże rozwój dziecka. Podpowiadamy, jak zachecić dziecko do nauki matematyki.

2. **Gold Summary:** Ciepłe, letnie dni aż sie prosza o aktywna zabawe na dworze. Dlatego jeśli planujesz przyjecie dla dziecka i jego przyjaciół, koniecznie postaw na garden party. Tego typu imprezy sa nie tylko modne, ale także łatwe w przygotowaniu i posprzataniu po nich. Jeśli nie masz pomysłu, jak je przygotować, podpowiadamy.

   plT5-base: Ozdabianie ogrodu to nie tylko świetna zabawa, ale także świetna rozrywka. Podpowiadamy, jak przygotować ogrodowe przyjecie dla dziecka.

   plBART: Dzieci uwielbiaja spedzać czas na świeżym powietrzu. To doskonała okazja do zorganizowania im urodzinowego przyjecia, które zachwyci ich nie tylko wygladem, ale także kreatywnościa i zapewni wiele pozytywnych emocji. Podpowiadamy, jak przygotować party dla maluchów w cenie do 50 zł.

   papuGaPT2: Organizujesz przyjecie urodzinowe dla dziecka? Zastanawiasz sie, co mu kupić, aby było mu jak najlepiej. Podpowiadamy, na co zwrócić uwage, wybierajac dekoracje do ogrodu. Sprawdź nasze propozycje.

   mT5-base: Ogrodowe przyjecie to wyjatkowe wydarzenie w życiu każdego dziecka. To świetna okazja do spedzenia czasu na świeżym powietrzu. Warto wiec zadbać o to, aby było przyjemne i kolorowe. Podpowiadamy, jakie dekoracje beda najlepsze dla twojego malucha.

   plT5-large Ozdabianie ogrodu to nie tylko świetna zabawa, ale także świetna rozrywka. Podpowiadamy, jak przygotować ogrodowe przyjecie dla dziecka.

F. Example predictions on PSC (whole)

1. **Gold Summary:** Ostatnia znaczna podwyżka stóp procentowych NBP oraz gwałtowne przyspieszenie inflacji wywołały krytyke poczynań Rady Polityki Pieniężnej. Rada chciałaby podjać dyskusje z głosami krytycznymi. nikt nie przewidywał, że w 1999 r. inflacja przewyższy cel inflacyjny NBP.Świadczy to dobitnie o tym, że głównymi czynnikami przyspieszajacymi inflacje były szoki podażowe na rynku żywności i paliw. Inny komentator stawia z kolei zarzut, że Rada niepotrzebnie zwlekała z decyzja o podwyżce stóp do listopada. Rada listopadowa decyzje nie dlatego podjeła w listopadzie, i w takiej skali, że "zaspała" w październiku, ale dlatego, by podjać ja 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. Pamietać jednak musi, że jego twierdzenie sprowadza sie do tezy: skala dokonanej podwyżki stóp procentowych doprowadzi do przestrzelenia celu inflacyjnego w dół. Czy racje maja ci, którzy krytykuja teraz Rade, dowiemy sie za rok. Przejdźmy do zarzutu zbyt dużej redukcji stóp procentowych 20 stycznia 1999 r. Otóż, redukcja rzeczywiście była zbyt duża badź 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 zasada forward looking. Czy Rada miała przesłanki do wcześniejszej zmiany swoich przewidywań? Bardzo mało i późno. o problemach ZUS dowiedzieliśmy sie dopiero w lipcu. Gwałtowny wzrost cen żywności rozpoczał sie 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 bedac odpowiedzialna za walke z inflacja, stara sie zrzucić wine na rzad i wydarzenia na rynku

miedzynarodowym, jest nierzetelne. Nie można winić straży pożarnej za lekkomyślność dziecka z zapałkami, jeśli okaże sie sprawna w gaszeniu pożaru. nie można ignorować krótkookresowego wpływu na inflacje czynników pozapieniężnych. nie da sie 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 krytyke poczynań Rady Polityki Pieniężnej. niestety, trudno podjać merytoryczna dyskusje z głosami, szczególnie w sytuacji niepowodzeń.Rada listopadowa decyzje podejmowała w całkowitym spokoju, bez żadnej nerwowości.Pieda zwracać uwage na to, że polityka pieniężna była zbyt pożadanym, zważywszy role, jaka odgrywaja. W tym kontekście stawianie Radzie zarzutów można było doprowadzić do końca września.

plT5-small: Ostatnia podwyżka stóp procentowych NBP oraz gwałtowne przyspieszenie inflacji wywołały krytyke poczynań Rady Polityki Pieniężnej. Rada listopadowa decyzje 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 badź w ogóle niepotrzebna.

papuGaPT2: redukcja stóp w styczniu była zbyt duża badź 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 dostepne były kilka tygodni po ich wystapieniu. niska przewidywalność polityki gospodarczej rzadu widać było gołym okiem. decyzja taka jest zrozumiała. redukcja stóp była koniecznym posunieciem, aby zbliżyć sie do rynku. jednak Rada przyjmowała wówczas za wiarygodna zapowiedź zaostrzenia polityki fiskalnej oraz przewidywała powolny wzrost eksportu. starano sie reanimować kryzys. pojawiła sie propozycja, żeby Rada zaczeła mówić jednym głosem. Dziś grono to znacznie sie powiekszyło. z kolejnych wypowiedzi zaczeli wycofać wniosek o rewizje nadzwyczajna. Rada listopadowa decyzje 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 publicznego. Ryzyko destabilizacji rynku jest wysokie. nie wolno zrezygnować z powrotu na przyzwoita ścieżke wzrostu ani z poprawienia równowagi zewnetrznej. Trzeba bardziej powściagliwie określić możliwości wycofania sie inflacji z polityki monetarnej. jeśli spojrzymy na deprecjacje parytetu, okazuje sie, ze zestrzelenie nastapiło nie z inicjatywy samego ROR-u, lecz z checi ponownego ożywienia poprzez odpartyjnienie poprzez podjecie działań, których w tej sytuacji jeszcze nie jesteśmy w stanie podjać.

mT5-base: Nie da sie na trwałe zmniejszyć inflacje 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 krytyke poczynań Rady Polityki Pieniężnej. Niestety, trudno podjać merytoryczna dyskusje z głosami, szczególnie w sytuacji niepowodzeń w ograniczaniu inflacja.

plBART: Ostatnia znaczna podwyżka stóp procentowych NBP oraz gwałtowne przyspieszenie inflacji wywołały ze strony niektórych komentatorów krytyke poczynań Rady Polityki Pieniężnej. Rada listopadowa decyzje 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 badź niepotrzebna. o tym wiemy dopiero teraz. stawianie Radzie zarzutu, że bedac odpowiedzialna za walke z inflacja, stara sie zrzucić wine na rzad, jest nierzetelne.

plT5-base: Rzad Polityki Pieniężnej w poczuciu odpowiedzialności przywiazuje wielkie znaczenie do przejrzystości swoich działań od poczatku istnienia. Dlatego chciałaby podjać dyskusje z głosami krytycznymi, szczególnie w sytuacji niepowodzeń w ograniczaniu inflacji. Znaczna cześć ostatnich krytycznych wypowiedzi wobec Rady sprowadzała sie właściwie do trzech zarzutów: o to, że ostatnia podwyżka stóp procentowych była ”spóźniona, zbyt nerwowa i przesadna” reakcja Rady, której ”optymizm co do inflacji trwał zbyt długo”, i która ”jakby próbowała nadrobić zaległości w walce z inflacja”.

2. **Gold Summary:** Dwaj słynni bokserzy zawodowi Amerykanin Evander Holyfield i Brytyjczyk Lennox Lewis szykuja sie do rewanżowego pojedynku. Ich pierwszy pojedynek zakończył sie, haniebnym według fachowców, remisem. Wygrany bedzie królem boksu zawodowego.Historia boksu jest pełna ekscytujacych rewanży, maja je również na swoim koncie Holyfield i Lewis.Obecny faworyt Lewis, jest aktualnym mistrzem świata WBC. Poczatkowy faworyt Holyfield, ma dwa pasy mistrzowskie WBA i IBF. Jeśli przegra

Holyfield przegra tym samym Don King, najsłynniejszego promotora boksu zawodowego.Walke transmituje Canal Plus.

mT5-small: W najbliższa sobote w Las Vegas dojdzie do kolejnego ”rewanżu stulecia”. Zmierza sie w nim dwaj najwybitniejsi obecnie bokserzy wagi cieżkiej, Amerykanin Evander Holyfield i Brytyjczyk Lennox Lewis. Osiem miesiecy temu, 13 marca, w nowojorskiej Madison Square Garden ich pierwsza walka zakończyła sie kontrowersyjnym remisem. Stawka rewanżowego pojedynku beda pasy mistrzowskie trzech najważniejszych organizacji boksu zawodowego: WBA, WBC, IBF.

plT5-small: W najbliższa sobote w Las Vegas dojdzie do kolejnego ”rewanżu stulecia”. Zmierza sie w nim dwaj najwybitniejsi obecnie bokserzy wagi cieżkiej, Amerykanin Evander Holyfield i Brytyjczyk Lennox Lewis. Osiem miesiecy temu, 13 marca, w nowojorskiej Madison Square Garden ich pierwsza walka zakończyła sie kontrowersyjnym remisem. Stawka rewanżowego pojedynku beda pasy mistrzowskie trzech najważniejszych organizacji boksu zawodowego: WBA, WBC, IBF. Ostatnim posiadaczem trzech pasów w wadze cieżkiej był Riddick Bowe. Od siedmiu lat żaden z pieściarzy nie może sie uważać za króla tej najbardziej cenionej kategorii.

papuGaPT2: W nocy z 13 na 14 listopada w Thomas and Mack Center w ”Ready World Press” zmierza sie dwaj znakomici zawodnicy. Ich walke bedzie można ogladać na żywo w Canal Plus od godz. 3:25. Retransmisja w niedziele bedzie miała miejsce już raz trzeci w historii walki, która miała być najwiekszym wydarzeniem ostatnich lat. Transmitowany przez Jamesa ”Bustera” pojedynek Sonny’ego Listona z Bobem Fitzsimmonsem w 1899 r. zakończył sie dyskwalifikacja jednego z zawodników. Walka okazała sie skandalem. Trzy niezależne postepowania wyjaśniajace, które prowadzono po tej walce, niewiele dały. Nikomu nie udowodniono winy. \nBoks jest tak stary jak świat. Już Homer opiewał walki na gołe pieści. Wielkim zwolennikiem walk bokserów był rzymski cesarz Kaligula. W 1350 roku, gdy naprzeciw siebie staneli Jeffries i O’Donnnel, dwóch zawodników zadało przeciwnikowi dwa razy wiecej ciosów, niż spalonych. Dwudziestotrzyletni wówczas Buster trafił prawym prostym. Teraz, na dziesieć metrów, już w pierwszej rundzie Floyda Pattersona spotkał sie z Joem Frazierem, Muhammadem Alim, Tunneyem i Michaelem Moorerem. O 3.00 rano, kiedy stoczyli już swoje ostatnie rugi, przyszedł czas na jeszcze jedna walke z ”Milionerem”, ”Drakula” z Jackiem Dempseyem.

mT5-base: W najbliższa sobote w Las Vegas dojdzie do kolejnego ”rewanżu stulecia”. Zmierza sie w nim dwaj najwybitniejsi obecnie bokserzy wagi cieżkiej, Amerykanin Evander Holyfield i Brytyjczyk Lennox Lewis. Osiem miesiecy temu, 13 marca, w nowojorskiej Madison Square Garden ich pierwsza walka zakończyła sie kontrowersyjnym remisem. Stawka rewanżowego pojedynku beda pasy mistrzowskie trzech najważniejszych organizacji boksu zawodowego: WBA, WBC, IBF.

plBART: W najbliższa sobote w Las Vegas dojdzie do kolejnego ”rewanżu stulecia”. Zmierza sie w nim dwaj najwybitniejsi obecnie bokserzy wagi cieżkiej, Amerykanin Evander Holyfield i Brytyjczyk Lennox Lewis. Osiem miesiecy temu, 13 marca, w nowojorskiej Madison Square Garden ich pierwsza walka zakończyła sie kontrowersyjnym remisem. Stawka rewanżowego pojedynku beda pasy mistrzowskie trzech najważniejszych organizacji boksu zawodowego: WBA, WBC, IBF. Ostatnim posiadaczem trzech pasów w wadze cieżkiej był Riddick Bowe. Od siedmiu lat żaden z pieściarzy nie może sie uważać za króla tej najbardziej cenionej kategorii.

plT5-base: W najbliższa sobote w Las Vegas dojdzie do kolejnego ”rewanżu stulecia”. Zmierza sie w nim dwaj najwybitniejsi obecnie bokserzy wagi cieżkiej, Amerykanin Evander Holyfield i Brytyjczyk Lennox Lewis. Osiem miesiecy temu, 13 marca, w nowojorskiej Madison Square Garden ich pierwsza walka zakończyła sie kontrowersyjnym remisem. Stawka rewanżowego pojedynku beda pasy mistrzowskie trzech najważniejszych organizacji boksu zawodowego: WBA, WBC, IBF. Ostatnim posiadaczem trzech pasów w wadze cieżkiej był Riddick Bowe. Od siedmiu lat żaden z pieściarzy nie może sie uważać za króla tej najbardziej cenionej kategorii.

G. Example predictions on PSC (extract)

1. **Gold Summary:** Politycy z pierwszych stron gazet szykuja sie do objecia stanowisk prezydentów dużych miast. Dla nich gotowi sa nawet zrezygnować z mandatów poselskich.\nW najbliższych tygodniach rozstrzygnie sie, czy w jesiennych wyborach samorzadowych prezydenci miast wybierani beda bezpośrednio przez mieszkańców. W Sejmie trwaja 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 sie Ryszarda Kalisza i Danute Waniek. ”Nie wyklucza” kandydowania w Warszawie Lech Kaczyński z Prawa i Sprawiedliwości. Trzecim liczacym sie kandydatem może być Andrzej Olechowski, lider Platformy Obywatelskiej. Nie stanie w szranki natomiast były prezydent Paweł Piskorski.

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

plT5-small: Politycy szykuja sie do objecia stanowisk prezydentów dużych miast. Dla nich gotowi sa nawet zrezygnować z mandatów poselskich. W najbliższych tygodniach rozstrzygnie sie, czy w jesiennych wyborach samorzadowych prezydenci miast wybierani beda bezpośrednio przez mieszkańców.

papuGaPT2: Partia ma poważne problemy z przywództwem. Nie ma mocnego nazwiska. Na razie próbuje sie rozwiazywać kolejne oświecimskie 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, chca go popierać.Próba stworzenia drugiego centrum politycznego nie powiodła sie. 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 Urzedzie Miasta.

mT5-base: W najbliższych tygodniach rozstrzygnie sie, czy w jesiennych wyborach samorzadowych prezydenci miast wybierani beda bezpośrednio przez mieszkańców. Tymczasem konkurenci siegaja 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 sie Ryszarda Kalisza i Danute Waniek. Te dwie osoby maja w stolicy duże poparcie.

plBART: Politycy szykuja sie do objecia stanowisk prezydentów dużych miast. Dla nich gotowi sa nawet zrezygnować z mandatów poselskich.W najbliższych tygodniach rozstrzygnie sie, czy w jesiennych wyborach samorzadowych prezydenci miast wybierani beda bezpośrednio przez mieszkańców. Jako kandydatów SLD na stanowisko w stołecznym ratuszu wymienia sie Ryszarda Kalisza i Danute Waniek. "Nie wyklucza" kandydowania w Warszawie Lech Kaczyński z Prawa i Sprawiedliwości. Trzecim liczacym sie kandydatem może być Andrzej Olechowski, lider Platformy Obywatelskiej.

plT5-base: Czy w jesiennych wyborach samorzadowych prezydenci miast wybierani beda bezpośrednio przez mieszkańców. W Sejmie trwaja prace nad projektem ustawy o bezpośrednim wyborze wójtów, burmistrzów i prezydentów. Ustawa ma poparcie głównych sił politycznych, jak SLD i PO, jest wiec wielce prawdopodobne, iż zostanie uchwalona. Nic dziwnego, że już dzisiaj partie zastanawiaja sie, kim obsadzić najwyższe stanowiska w miastach. Dla niektórych polityków odejście do samorzadu oznaczałoby konieczność złożenia mandatu posła, ale nie zraża ich to.

2. **Gold Summary:** Ludzie całkowicie sparaliżowani beda mogli sama myśla sterować komputerem, by porozumieć sie z otoczeniem. Niezwykle czułe detektory, odczytujace wybrane fale mózgowe, testowane sa w laboratoriach USA, Japonii i Europy. Do łaczności z urzadzeniami elektronicznymi można wykorzystać jakakolwiek aktywność układu nerwowego - sygnały elektryczne wytwarzane ruchami gałek ocznych lub impulsy nerwów mieśni.

   mT5-small: Uniwersytetu Emory w Atlancie wypróbowali u dwóch pacjentów system pozwalajacy sterować kursorem na ekranie monitora. Sterowanie komputerem sama myśla stało sie możliwe, gdy uczeni lepiej opanowali odczytywanie i werbalizacje sygnałów. Komunikowanie sie przy użyciu takiego systemu jest dość żmudne, ale wyłacznie dla ludzi zdrowych. Dla osób całkowicie sparaliżowanych, nie mogacych wykonywać żadnych ruchów, nawet mieśniami twarzy, jest jedynym sposobem porozumiewania.

   plT5-small: Ludzie całkowicie sparaliżowani beda mogli sama myśla sterować komputerem, by porozumieć sie z otoczeniem. Niezwykle czułe detektory, odczytujace wybrane fale mózgowe, testowane sa w laboratoriach USA, Japonii i Europy. Do łaczności z urzadzeniami elektronicznymi można wykorzystać jakakolwiek aktywność układu nerwowego - sygnały elektryczne wytwarzane ruchami gałek ocznych lub impulsy nerwów mieśni. Jeszcze niedawno wszelkie rozważania na ten temat były domena wyłacznie autorów powieści fantastycznonaukowych. Obecnie naukowcy w kilkunastu ośrodkach prześcigaja sie w konstruowaniu urzadzeń coraz sprawniej wykrywajacych komunikaty myślowe.

   papuGaPT2: dzisiaj coraz wiecej teoretyków na temat tego, co może sie stać, jak bedzie możliwe, za kilka lat. Trudno wyobrazić sobie wielka liczbe skomplikowana, krecacych sie po krzywej Kenii i Epsilon. Gdyby jednak wziać pod uwage normalny, a nie ten podkolorowany obraz powyborczej rzeczywistości, to rzeczywiście wygladałby nedzny. Jeszcze niedawno takie pytanie byłoby niedorzeczne. Dziś naukowcy nie sa już takimi sceptykami. \nNiektórzy jeszcze przed przystapieniem Polski do UE zastanawiali sie, czy w ogóle wchodzić w ten projekt. Nie wszyscy sa entuzjastami nowego tworu, który stał sie również przedmiotem marzeń i zazdrości milionów. A Polska? Jak wynika z sondaży, 77 proc. Amerykanów gotowych

jest poprzeć likwidacje lub gruntowna zmiane europejskich norm. Przeciwnicy wejścia do Unii Europejskiej obawiaja sie takiego myślenia, bo bedzie 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 domena wyłacznie autorów powieści fantastycznonaukowych. Obecnie uczeni w kilkunastu ośrodkach prześcigaja sie w konstruowaniu urzadzeń coraz sprawniej wykrywajacych komunikaty myślowe. Niezwykle czułe detektory, odczytujace wybrane fale mózgowe, testowane sa w laboratoriach USA, Japonii i Europy. Kilka takich rozwiazań jest testowanych w USA.

plBART: Ludzie całkowicie sparaliżowani beda mogli sama myśla sterować komputerem, by porozumieć sie z otoczeniem. Niezwykle czułe detektory, odczytujace wybrane fale mózgowe, testowane sa w laboratoriach USA, Japonii i Europy. Do łaczności z urzadzeniami elektronicznymi można wykorzystać jakakolwiek aktywność układu nerwowego. Urzadzenie ma jednak te wade, że wymaga użycia wszczepów, grożacych powstaniem infekcji i uszkodzeniem mózgu.

plT5-base: ZBIGNIEW WOJTASIŃSKI Ludzie całkowicie sparaliżowani beda mogli sama myśla sterować komputerem, by porozumieć sie z otoczeniem. Niezwykle czułe detektory, odczytujace wybrane fale mózgowe, testowane sa w laboratoriach USA, Japonii i Europy. Do łaczności z urzadzeniami elektronicznymi można wykorzystać jakakolwiek aktywność układu nerwowego - sygnały elektryczne wytwarzane ruchami gałek ocznych lub impulsy nerwów mieśni.

H. Example predictions on PSC (abstract)

1. **Gold Summary:** Anachroniczna jest koncepcja, u nas bynajmniej nierzadko wyznawana, że swoboda umów wyklucza idee ochrony konsumenta.Przed kilkoma laty niemiecki Trybunał Konstytucyjny uznał, że nawet osoba pełnoletnia, samodzielna, nie poddana żadnemu przymusowi wymaga ochrony reczac za kredyt bankowy oraz że konieczna jest szczegółowa, wyczerpujaca, niedwuznaczna, jasna informacja, ze wskazaniem na kwestie najbardziej niebezpieczne dla poreczyciela. Lepsza informacja dla konsumenta łagodzi bowiem nierówność pozycji rynkowej. W konsekwencji tych orzeczeń zmieniła sie praktyka powszechnych sadów w Niemczech. Trzy kwestie zasługuja tu na uwage. Po pierwsze - umowy kredytowe i sytuacja poreczyciela doczekały sie w Niemczech oceny TK, dokonywanej z konstytucyjnego punktu widzenia. Po drugie - TK za remedium na strukturalne zachwianie równowagi umownej uznał zwiekszenie obowiazków 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 idee ochrony konsumenta.Z konstytucji nie można wyczytać, jakimi środkami i na jakim poziomie ma sie chronić konsumenta, ale to nie jest jedyny możliwy sposób "użycia konstytucji" do takich celów. Ustawa zasadnicza da sie użyć jako norma rozstrzygajaca 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 sad musi odkodować, nadać im konkretna treść.Druga ważna kwestia jest znaczenie informacji jako oreża konsumenta. W europejskim prawie wspólnotowym kamieniem wegielnym ochrony konsumenta jest informacja. Uważa sie, że konsument wymaga ochrony, ponieważ jest źle poinformowany i na skutek tego nie może w prawdziwie wolny i nieskrepowany sposób decydować o swym "udziale na rynku". Stad sie biora niezwykle rozbudowane w dyrektywach i ich implementacjach wewnatrzkrajowych przepisy mówiace, o czym, kiedy i jak trzeba konsumenta informować. Nie tylko jednak sama obrona przez informacje jest cecha charakterystyczna europejskiego prawa konsumenckiego. Wystepuje tu jeszcze wskazanie minimalnego poziomu treści umowy. Oznacza to, że poszczególne państwa, implementujac dyrektywy konsumenckie we własnym porzadku prawnym, nie moga zejść niżej wskazanego tam poziomu ochrony. Ale tutaj tkwi pułapka: przyznanie w prawie krajowym zbyt wysokiego poziomu ochrony, może być uznane za rzeczywista czy ukryta praktyke dyskryminacyjna, ograniczajaca dostep towarów i usług na rynek wspólny. Nie tylko informacja i nie tylko określenie minimalnego poziomu uprawnień zagwarantowanych konsumentowi w dotyczacym danej umowy prawie wspólnotowym czy wewnetrznym sa cechami charakterystycznymi europejskiego prawa konsumenckiego. Otóż cecha charakterystyczna współczesnego prawa umów jest to, że konsumentowi przysługuje prawo "wycofania sie z transakcji". W praktyce i u nas ta nowa technika ochronna zaczyna sie pojawiać: zna ja polskie prawo ubezpieczeniowe. Co charakterystyczne: tego rodzaju możliwość wycofania sie z transakcji służy (w prawie europejskim) jedynie konsumentowi, nie jego kontrahentowi. I znów jest to jeden ze środków, który przyznaje sie tylko stronie instytucjonalnie słabszej, używany, by zrównoważyć siłe rynkowa 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 idee ochrony konsumenta. Po trzecie - umowy kredytowe i sytuacja poreczyciela doczekały sie w Niemczech oceny TK, dokonywanej z konstytucyjnym punktu widzenia.

plT5-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 reczac za kredyt bankowy. Druga istotna kwestia europejskiego prawa konsumenckiego jest wskazanie minimalnej wiedzy i doświadczenia miedzy poreczycielem i bankiem. Innym przykładem środka ochrony konsumenta jest niemiecki Trybunał Konstytucyjny. Moim zdaniem, konsumentowi przysługuje prawo wycofania sie 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 wskazuja minimalny poziom treści umowy: sa to przepisy zakazujace 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 ochrone przed wprowadzajacymi w błąd informacjami na etykietach. Wreszcie cecha charakterystyczna współczesnego prawa umów jest to, iż konsument przez krótki czas (siedem do dziesieciu dni) ma prawo, bez żadnych negatywnych konsekwencji, zrezygnować z transakcji. Środek ten przysługuje tylko konsumentowi jako stronie instytucjonalnie słabszej; jego kontrahenta obowiazuje ogólna zasada mówiaca o ”umowy powinny zostać dotrzymywane”. Zasada ta wyraża sie 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 siłe przyciagania i działać na rzecz zrekompensowania braków jego wiedzy. Stad sie biora rozbudowane przepisy mówiace, o czym, kiedy i jak należy informować konsumenta o wszelkich aspektach umowy kredytowej, a także o możliwych niebezpieczeństwach i ryzykach, które moga być nieświadomym wyrazem orientacji konsumenta w transakcjach rynkowych.

mT5-base: Re swoboda umów wyklucza idee 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 reczac za kredyt bankowy, bo nierównowaga wiedzy i doświadczenia miedzy poreczycielem i bankiem jest zbyt wielka i tolerowanie jej zagraża jednej z konstytucyjnych zasad niemieckiego porzadku prawnego: wyrażonej w art. 2 konstytucji autonomii woli. W konsekwencji orzeczeń TK zmieniła sie praktyka niemieckich sadów, które umowy kredytowe doczekały sie 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 idee ochrony konsumenta.

plT5-base: W.

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

   mT5-small: Nakładem wydawnictwa Huberta Orłowskiego poświecona jest stereotypom dotyczacym stosunków polsko-niemieckich, ze szczególnym uwzglednieniem najsilniejszych z nich: polnische Wirtschaft”. Autorstwo tego stereotypu przypisuje sie niejakiemu Georgowi Fosterowi, uczonemu i podróżnikowi z Żuław Wiślanych. ”Polnische Wirtschaft” oznaczajacemu kombinacje tych cech. Wśród warstw bogatszych widziano rozrzutność, równocześnie jednak także brak higieny, zacofanie, ignorancja.

   plT5-small: W ksiażce profesora Huberta Orłowskiego poświecona jest stereotypom dotyczacym stosunków polsko-niemieckich, ze szczególnym uwzglednieniem najsilniejszych z nich: ”polnische Wirtschaft” oznaczajacemu kombinacje bałaganu, nieudolności, brudu i zła gospodarke.

   papuGaPT2: Od lat 40. literatura podtrzymywała mit szlachetnego Polaka, o którym zapomina młodsze pokolenie. Stereotyp ten nigdy by może nie zaistniał, gdyby nie głebokie różnice dzielace polskie cnoty szlacheckie od niemieckich cnót mieszczańskich, pogłebione różnicami miedzy katolicyzmem i protestantyzmem. Po klesce powstania listopadowego, po upadku muru berlińskiego okazało sie, że ustrój ten

nie był w stanie zdobyć uznania i poparcia ludu, dlatego też w ksiażkach, operetkach i filmach pełno jest dowcipów o polskiej gospodarce i polskim sejmikowaniu. Profesor Hubert Orłowski jest współautorem tzw. ustawy teczkowej, ustanawiajacej Instytut Pamieci Narodowej, oraz ustawy o ochronie informacji niejawnych, określajacej m.in. mianem ”pacta servanda sunt” na której spoczywa cała odpowiedzialność.

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

plBART: Praca profesora Huberta Orłowskiego ”Polnische Wirtschaft. Nowoczesny niemiecki dyskurs o Polsce” poświecona jest stereotypom dotyczacym stosunków polsko-niemieckich, ze szczególnym uwzględnieniem najsilniejszych z nich: polnischer Wirthschaft”. Termin ten określa ”nieporzadek i zła gospodarke, niedbalstwo i bałagan”. Autorstwo tego, majacego długa historie i niesłychana żywotność, stereotypu przypisuje sie niejakiemu Georgowi Fosterowi, uczonemu i podróżnikowi z Żuław Wiślanych. Naprawde jednak formuła była już gotowa i nieźle znana wcześniej, choć w myśleniu Niemców (i nie tylko ich) zadomowiła sie 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 przywiazaniu do rodzinnych zwyczajów nie chce pisać już nic wiecej, aby nie przedłużać tego listu”.

plT5-base: W.