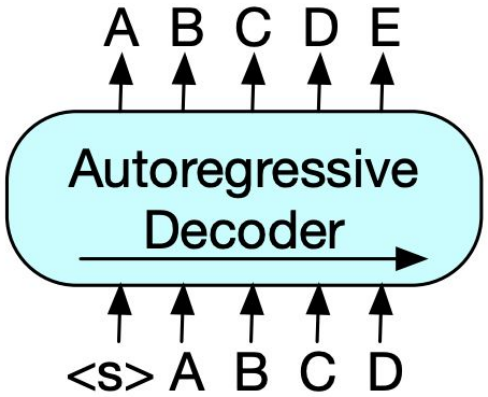
Evaluation of Transfer Learning for Polish with a Text-to-Text Model

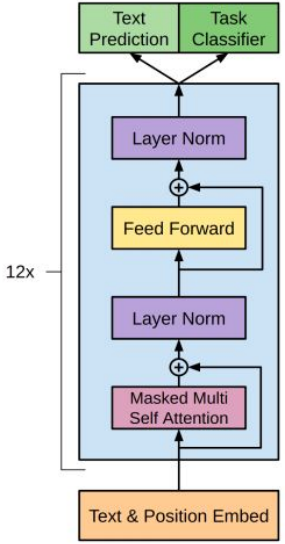
Aleksandra Chrabrowa, Łukasz Dragan, Karol Grzegorczyk, Dariusz Kajtoch
Mikołaj Koszowski, Robert Mroczkowski, Piotr Rybak

allegro

1. Available text-to-text models
2. KLEJ benchmark results
3. Machine translation task
4. Q&A task
5. Summarization task

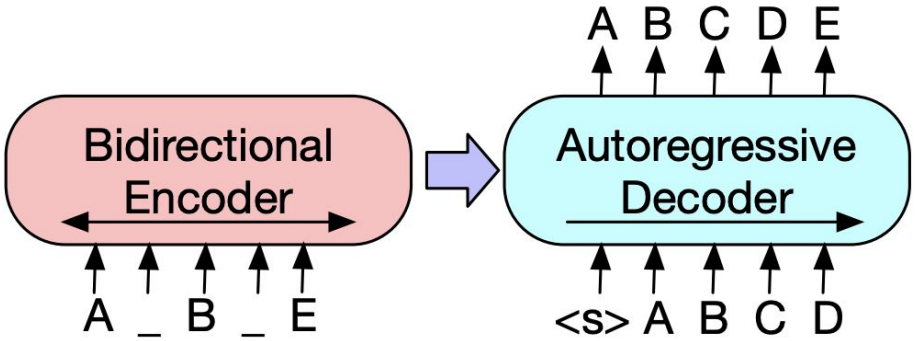
The GPT-2 model (papuGaPT2)





Alec Radford, et. al., *Language Models are Unsupervised Multitask Learners*, 2018

The BART model (plBART)



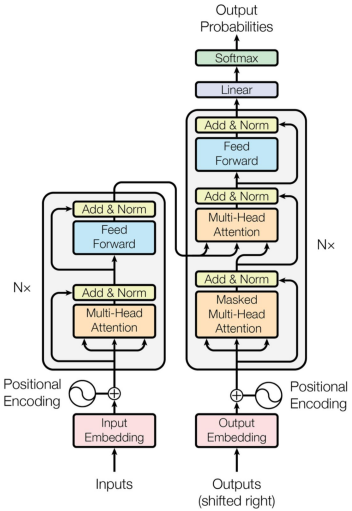
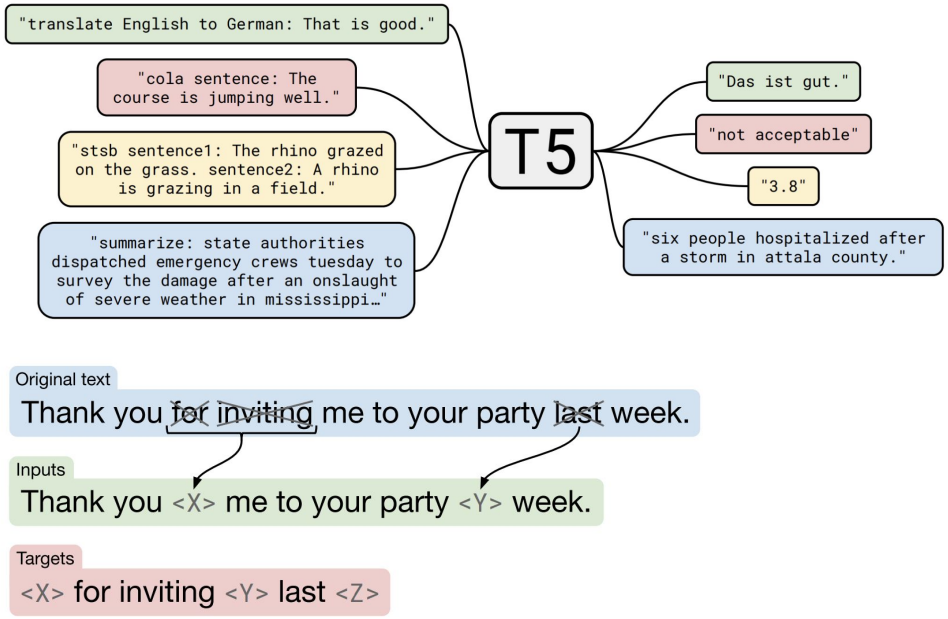


Figure 1: The Transformer - model architecture.

Mike Lewis, et. al., *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*, ACL 2020

The T5 model



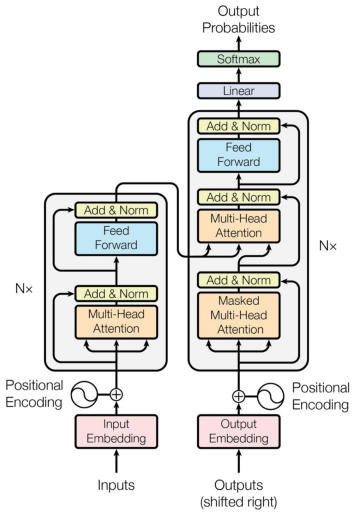


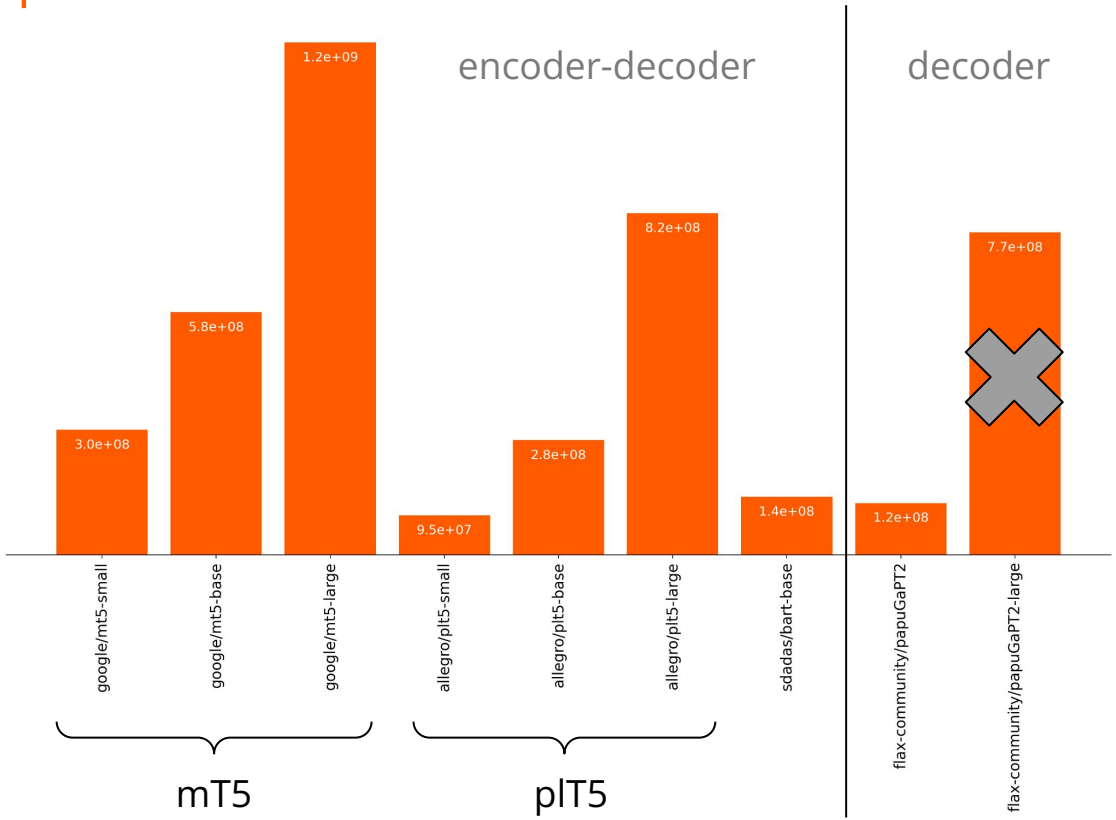
Figure 1: The Transformer - model architecture.

Colin Raffel, et. al., *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*, JMLR 2020
Linting Xue, et. al., *mT5: A massively multilingual pre-trained text-to-text transformer*, ACL 2021

The Polish T5 model (plT5)

- Initialized from mT5 checkpoint
- Embedding layer was shrinked from 250k to 50k tokens (see HerBERT paper)
- Trained for 50k steps on a single TPU v3
- Dataset mixture of wikipedia, wolne lektury, nkjp, open subtitles, CCNet
- Checkpoints publicly available from Transformers Hub (small/base/large)

Trainable parameters



KLEJ Benchmark

allegro

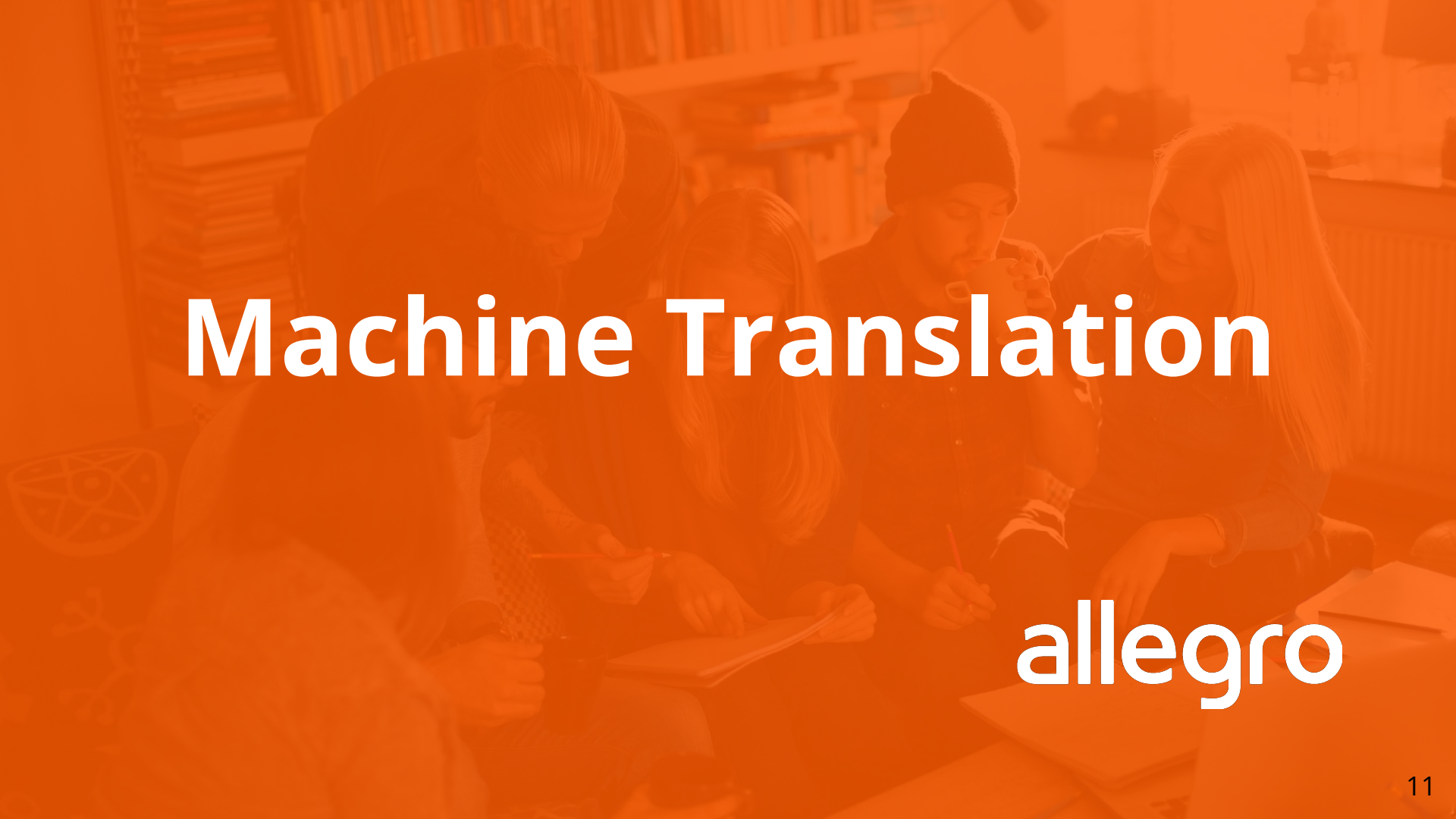
Task construction

- 7 different NLU tasks

Label is generated greedily

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

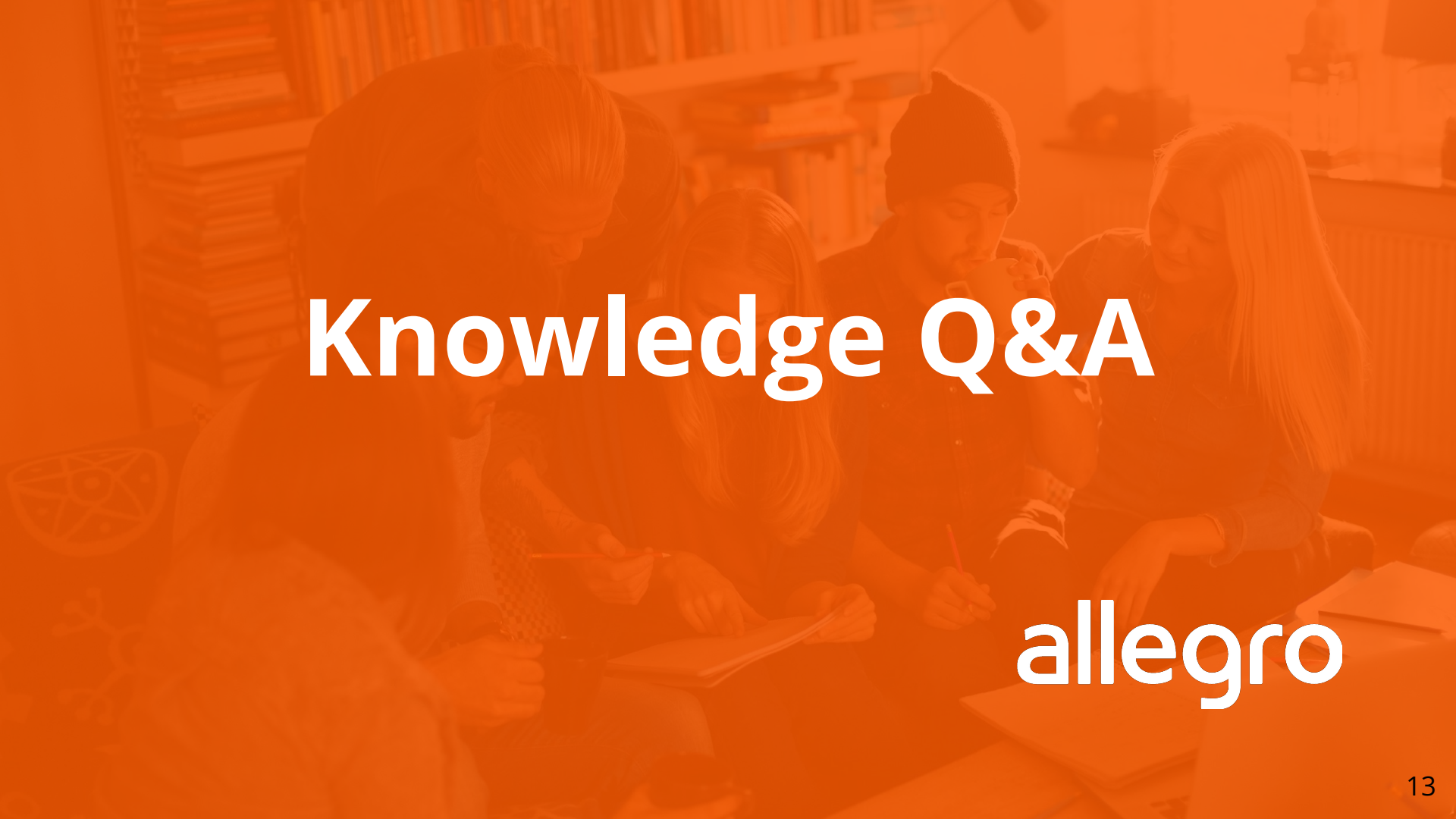
- The larger the model the better the results
- Text-to-text models perform worse than HerBERT
- BART is performing extremely well even though it does not have as many parameters as base plT5
- GPT-2 is not bad, but struggles with *CBD* and *Czy wiesz?* tasks
- plT5 has best performance


Machine Translation
allegro

Model		Vocab	en→pl	pl→en
Small models	mT5	mT5	20.5	25.0
	plT5	plT5	20.3	24.7
	mT5	wmt20	20.8	24.8
	plT5		20.8	25.4
Base models	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

- The larger the model the better the results
- Translation to Polish language has better performance with plT5 (polish vocabulary)
- Task specific tokenizer (wmt20) improves performance (although not so clear for plT5)
- Reproduced known fact that translation to Polish has lower absolute BLEU score than to English
- T5 models give best results

WMT20 - task-specific vocab with 32k tokens


Knowledge Q&A
allegro

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**

- In general all text-to-text models struggle with knowledge Q&A (without passage)
- All models have performance above baseline
- Pretraining on Wikipedia helps
- We observed that BART reached final performance much faster than plT5
- plT5 gives best results

Summarization

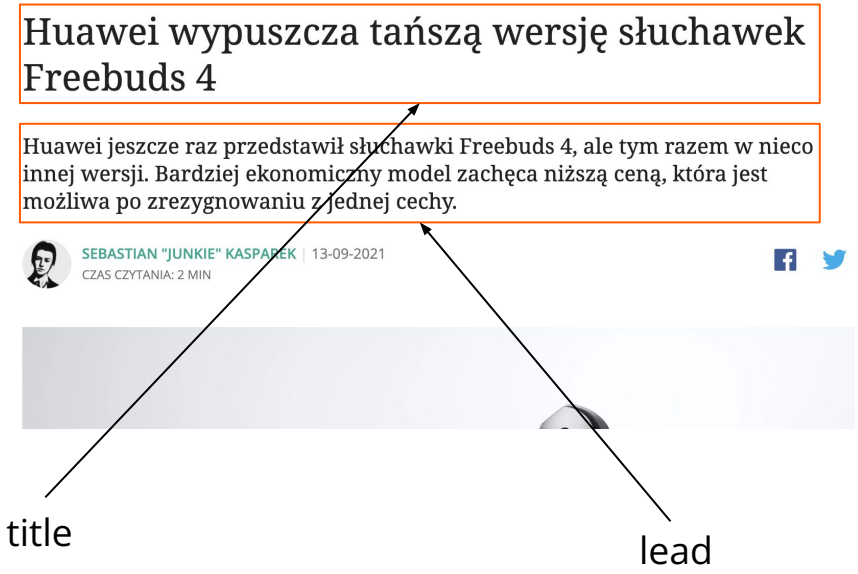
allegro

Datasets

Name	Subset	Train	Dev	Test	Domain
AA	whole	24363	2708	6768	E-commerce articles
PSC	whole	7799	867	2167	News articles
	extract	6137	682	1705	
	abstract	1662	185	462	

AA - Allegro Articles (https://allegro.pl/artykuly)
PSC - Polish Summaries Corpus (https://huggingface.co/datasets/psc)

Datasets (Allegro Articles)



Słuchawki Huawei Freebuds 4 doczekały się nowej, lekko uboższej wersji. Bardziej ekonomiczna propozycja nie różni się znacząco od modelu wypuszczonego w maju bieżącego roku.

Bez bezprzewodowego ładowania, ale za to taniej

Huawei, gigant przemysłu smartfonów, zdecydował się wprowadzić na rynek chiński rewizję słuchawek Freebuds 4. Nowa wersja rządzenia miała już swoją premierę i została pozbawiona funkcji bezprzewodowego ładowania. **Odbiło się to korzystnie na cenie, która spadła z poziomu 999 juanów (około 596 zł) do 899 juanów (około 537 zł)**.

Zrezygnowanie z bezprzewodowego ładowania to jedyna różnica względem pierwotnej wersji Freebuds 4. Reszta specyfikacji pozostała bez zmian. Słuchawki działają na układzie Kirin A4, mają przetworniki dynamiczne 14,3 mm i wspierają funkcję aktywnej redukcji szumów oraz Bluetooth w wersji 5.2. Odporne na zachlapania urządzenie wyposażono w akumulator, który starcza na 4 godzin działania, a dołączone etui potrafi wydłużyć ten czas do 22 godzin.

body

https://allegro.pl/artykul/huawei-wypuszcza-tansza-wersje-sluchawek-freebuds-4-VLEWE8Da9uj

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

- Polish BART has best performance hypothetically due to “copy bias”
- Best model is usually **2-3pp** above baseline with exception of PSC abstract (below baseline) and AA body+lead2title (**32pp** above baseline)
- Not all models are able to generate summaries that are better than baseline
- Larger models are unstable and degrade performance
- Second best model is plT5

Reported metric: average of (f-measure) ROUGE-1, ROUGE-2 and ROUGE-L

Summary

allegro

- We trained T5 model for Polish language
- We benchmarked plT5, polish BART and PapuGaPT2 on various text-to-text tasks
- In general, the larger the model the better the results
- Text-to-text models are still below BERT performance on KLEJ benchmark
- plT5 performs best in en-pl machine translation
- Knowledge Q&A is very difficult for all models
- BART model is performing very well especially on news summarization due to “copy bias”

THANK YOU!

dariusz.kajtoch@allegro.pl

allegro