Zasoby i narzędzia do automatycznego odpowiadania na pytania (QA)

Ryszard Tuora Natalia Zawadzka-Paluektau Cezary Klamra
Aleksandra Zwierzchowska Łukasz Kobyliński

Instytut Podstaw Informatyki, Polska Akademia Nauk

12 maja 2023

Wstęp

Question Answering (**QA**) — Odpowiadanie na pytania

Pomysł: Chcemy odpytywać zasoby informatyczne korzystając z języka naturalnego, bez potrzeby tłumaczenia tego na języki formalne — w tym różnego rodzaju języki zapytań (*query*).

Jak można wykorzystać do tego narzędzia **NLP**?

Wstęp

Problem: Każde zadanie można sformułować jako pytanie, więc **QA** jest zadaniem „AI-complete”.

Heterogeniczność kategorii pytań:

- Gdzie urodził się Kopernik?
- Czy Kopernik urodził się na terenie Polski?
- Ile lat Kopernik spędził na Akademii Krakowskiej?
- Które dokonanie Kopernika jest najważniejsze?
- Jak udowodnić Twierdzenie Kopernika?

Retrieve and Read

Prosta architektura, składająca się z dwóch głównych elementów:

1. **Retriever:** Przeszukuje bazę dokumentów by zwrócić relewantne dokumenty
2. **Reader:** Na podstawie wyników wyszukiwania, syntetyzuje odpowiedź

Oba komponenty mogą zostać oparte na uczeniu maszynowym, potrzebują wówczas danych.

Istniejące zasoby

Dla języka polskiego istnieją:

1. *CzyWiesz?* [Marcińczuk et al., 2013], 5 tys. par pytanie — artykuł z Wikipedii
2. Dane w zbiorach wielojęzycznych: *MKQA* [Longpre et al., 2020], *XQA* [Liu et al., 2019] i *MFAQ* [Bruyn et al., 2021]
3. *PolQA* [Rybak et al., 2022], zbiór danych do retrievalu, 7 tys. pytań + kilkadziesiąt fragmentów-kandydatów
4. *MAUPQA* [Rybak, 2023], zbliżony zbiór 'weakly labeled'
5. Maszynowo przetłumaczony na polski[Borzymowski, 2020] zbiór *SQuAD*[Rajpurkar et al., 2016] .

SQuAD

SQuAD[Rajpurkar et al., 2016]

- 150 tys. pytań, sformułowanych do fragmentów z angielskiej Wikipedii
- Dla każdego z pytań, odpowiedź stanowi minimalny *span* w tekście

SQuAD 2.0[Rajpurkar et al., 2018] wprowadza kategorię *pytań nieodpowiadalnych*, i.e. takich, które są relewantne dla fragmentu, ale nie można na nie na jego podstawie odpowiedzieć.

[d'Hoffschmidt et al., 2020, Möller et al., 2021] wskazują, że dostarczenie natywnych danych poprawia wyniki modeli.
`https://rajpurkar.github.io/SQuAD-explorer/explore/v2.0/dev/Harvard_University.html`

Zbiory SQuADopodobne

Tabela: Porównanie zbiorów opartych na formule SQuAD

Lang.	Year	Dataset	#QA pairs	SOTA F1	Imp. Qs.
eng	2018	SQuAD[Rajpurkar et al., 2016, Rajpurkar et al., 2018]	151 k	93.2	+
fra	2022	FQuAD[d'Hoffschmidt et al., 2020, Heinrich et al., 2022]	80 k	92.2	+
deu	2021	GermanQuAD[Möller et al., 2021]	14 k	88.1	–
slk	2023	SK-QuAD[Hládek et al., 2023]	91 k	74.4	+
rus	2019	SberQuAD[Efimov et al., 2020]	90 k	84.8	–
fas	2022	PQuAD[Ayoubi, 2021]	80 k	87.6	+
kor	2019	KorQuAD[Lim et al., 2019]	70 k	90.2	–
jpn	2022	JaQuAD[So et al., 2022]	40 k	78.9	–
vie	2021	UIT-ViQuAD[Nguyen et al., 2020]	23 k	87.0	–

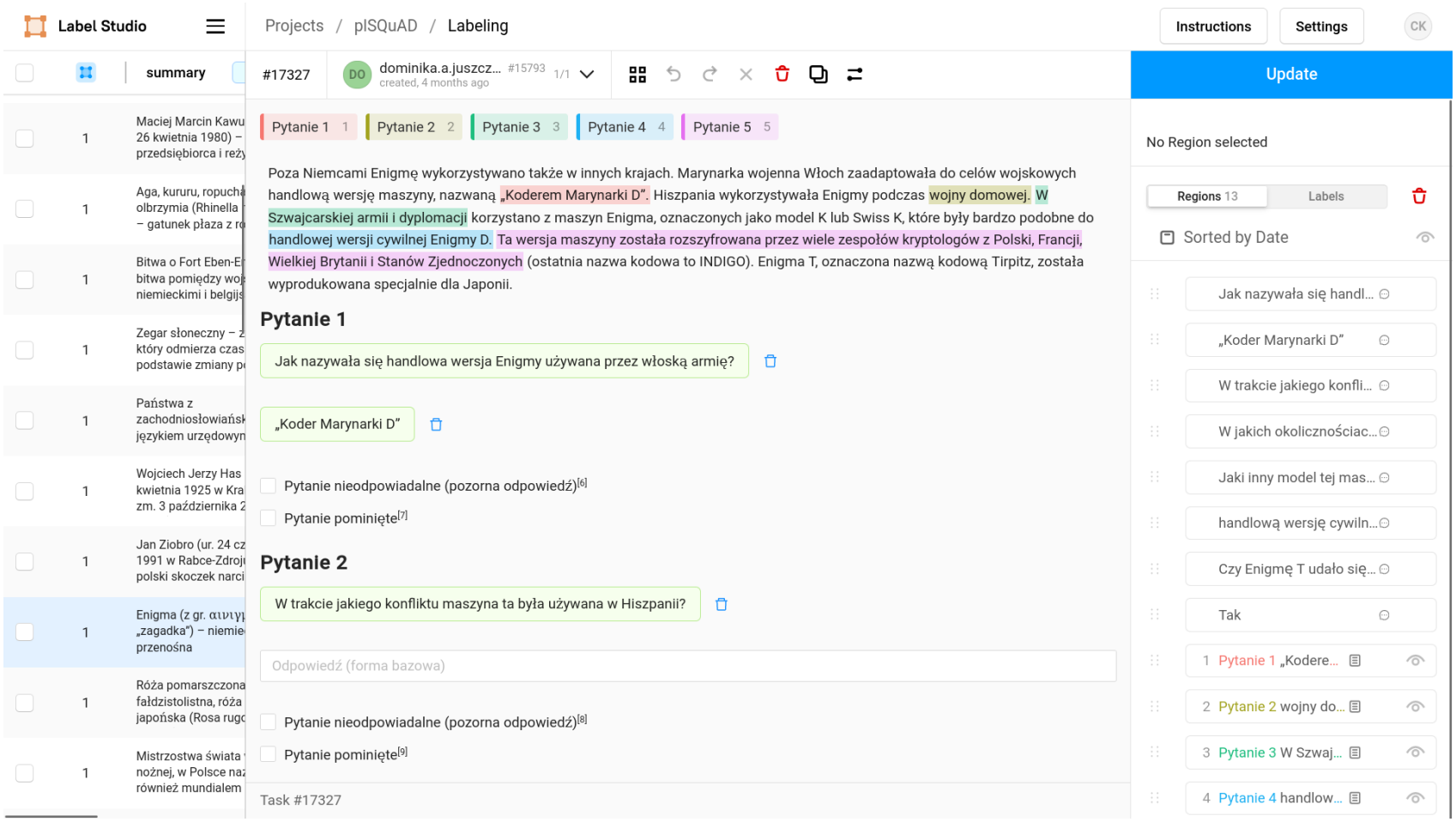
Zbieranie danych

PoQuAD — *The Polish Question Answering Dataset*

- 11 tys. artykułów z polskiej Wikipedii (m.in. na podstawie etykiety *Dobry Artykuł* oraz popularności)
- przeważnie jeden paragraf *per* artykuł
- wybór artykułów: centralność (mierzona `TextRank`) i długość ($<$ 500 znaków)
- 7 anotatorek, pracujących na podstawie instrukcji, ewaluacji i konsultacji
- zebrano 70,764 trójek pytanie-kontekst-odpowiedź
- train-dev-test-split 8:1:1

`https://huggingface.co/datasets/clarin-pl/poquad`

Interfejs


Label Studio
Projects / plSQuAD / Labeling
Instructions
Settings
CK
summary
#17327
dominika.a.juszcz... #15793
created, 4 months ago
1/1
Update
Pytanie 1 1
Pytanie 2 2
Pytanie 3 3
Pytanie 4 4
Pytanie 5 5
Poza Niemcami Enigmę wykorzystywano także w innych krajach. Marynarka wojenna Włoch zaadaptowała do celów wojskowych handlową wersję maszyny, nazwaną „Koderem Marynarki D". Hiszpania wykorzystywała Enigmy podczas wojny domowej. W Szwajcarskiej armii i dyplomacji korzystano z maszyn Enigma, oznaczonych jako model K lub Swiss K, które były bardzo podobne do handlowej wersji cywilnej Enigmy D. Ta wersja maszyny została rozszyfrowana przez wiele zespołów kryptologów z Polski, Francji, Wielkiej Brytanii i Stanów Zjednoczonych (ostatnia nazwa kodowa to INDIGO). Enigma T, oznaczona nazwą kodową Tirpitz, została wyprodukowana specjalnie dla Japonii.
Pytanie 1
Jak nazywała się handlowa wersja Enigmy używana przez włoską armię?
„Koder Marynarki D"
Pytanie nieodpowiadalne (pozorna odpowiedź)[6]
Pytanie pominięte[7]
Pytanie 2
W trakcie jakiego konfliktu maszyna ta była używana w Hiszpanii?
Odpowiedź (forma bazowa)
Pytanie nieodpowiadalne (pozorna odpowiedź)[8]
Pytanie pominięte[9]
Task #17327
No Region selected
Regions 13
Labels
Sorted by Date
Jak nazywała się handl...
„Koder Marynarki D"
W trakcie jakiego konfli...
W jakich okolicznościac...
Jaki inny model tej mas...
handlową wersję cywiln...
Czy Enigmę T udało się...
Tak
1 Pytanie 1 „Kodere...
2 Pytanie 2 wojny do...
3 Pytanie 3 W Szwaj...
4 Pytanie 4 handlow...

Warstwa Abstraktywna

Dodatkowa warstwa odpowiedzi *free-form*, służąca głównie uniezależnieniu od formy gramatycznej użytej w tekście (rozwiązanie zbliżone do [Sabol et al., 2019]). Odpowiedzi abstraktywne mogą powstawać w wyniku:

1. transformacji fleksyjnych
2. derywacji
3. dodania słów (np. przyimków)
4. usuwania słów (np. wtrąceń)

Q: W którym kraju, po raz pierwszy, wykryto obecność fosforiaku w warunkach naturalnych?
A: Na Węgrzech
Kontekst: [...] Po raz pierwszy jego obecność w gazach pochodzenia biologicznego stwierdzono w gazach wydobywających się z osadów w oczyszczalniach ścieków i płytkich jezior na terenie Węgier za pomocą spektrometrii mas w 1988 r., a następnie w wielu lokalizacjach na świecie. [...]

Pytania o rozstrzygnięcie

Dla pytań o rozstrzygnięce, w warstwie ekstraktywnej odpowiedzią jest najczęściej zdanie.

Q: Czy sąd wyższej instancji podtrzymał początkowy wyrok?
A: Nie.
Kontekst: "W marcu 2016 Breivik wytoczył proces przeciwko Norwegii, gdyż warunki w jakich jest przetrzymywany w więzieniu, uznał za „nieludzkie". Sąd częściowo przyznał mu rację, jednak norweski minister sprawiedliwości zapowiedział odwołanie od wyroku. 1 marca 2017 sąd apelacyjny w mieście Skien zmienił wyrok sądu niższej instancji i uznał, że warunki więzienne, w jakich przebywa Breivik, nie są nieludzkim i poniżającym traktowaniem. Sąd apelacyjny stwierdził, że „rygorystyczne środki są konieczne ze względów bezpieczeństwa" i uznał za właściwe odizolowanie skazanego od innych więźniów.",

Pytania nieodpowiadalne i pytania o rozstrzygnięcie

	odpowiadalne	nieodpowiadalne	Σ
o dopełnienie	49303	11787	61090
o rozstrzygnięcie	8451	1223	9674
Σ	57754	13010	70764

	odpowiadalne	nieodpowiadalne	Σ
o dopełnienie	69.67%	16.66%	86.33%
o rozstrzygnięcie	11.94%	1.73%	13.67%
Σ	81.61%	18.39%	100%

Tabela: Proporcje klas pytań

Statystyki pytań

Tabela: Distribution of question words (after lemmatization)

Question word	Freq. (%)
jaki/który	45.04
czy	14.09
co	8.48
kto/czyj	8.46
ile	6.29
jak	5.53
kiedy	4.89
gdzie/dokąd/skąd	3.60
dlaczego/czemu	3.58

Tabela: Answer entity type distribution

Ans. entity type	Freq. (%)
NONE	64.51
PERSNAME	10.32
NUMBER	7.05
DATE	6.07
PLACENAME	5.24
ORGNAME	4.46
GEOGNAME	1.38
MIXED	0.92
TIME	0.05

Tabela: Evaluation results on PoQuAD

QA paradigm	Train set	Model	HasAns		NoAns	Total	
			EM	F1	EM	EM	F1
extractive		human baseline	65.67	83.78	84.14	69.10	83.84
	PoQuAD	mBERT-base[Devlin et al., 2018]	52.52	68.45	52.14	52.46	65.49
		XLM-R-base[Conneau et al., 2019]	55.14	71.00	48.87	54.01	67.00
		XLM-R-large[Conneau et al., 2019]	58.04	74.44	57.91	58.02	71.44
		HerBERT-base[Mroczkowski et al., 2021]	59.95	75.82	54.40	58.95	71.94
		HerBERT-large[Mroczkowski et al., 2021]	64.52	80.56	**64.77**	**64.56**	**77.70**
	SQuAD-PL	mBERT-base	34.68	52.39	39.98	35.77	49.84
		XLM-R-base	37.44	54.88	41.00	38.17	52.03
		XLM-R-large	41.52	61.20	45.28	42.29	57.93
		HerBERT-base	44.70	64.36	36.17	42.95	58.57
		HerBERT-large	50.48	72.10	37.72	47.86	65.04
	SQuAD-PL + fine-tune on PoQuAD	HerBERT-large	**64.69**	**81.58**	59.24	63.70	77.54
abstractive		human baseline	64.62	81.15	84.14	68.25	81.71
	PoQuAD	mT5-base[Xue et al., 2020]	52.89	68.80	20.58	47.04	60.07
		BART-base[Dadas,]	48.61	66.31	28.99	45.06	59.55
		plT5-base[Chrabrowa et al., 2022]	58.30	73.53	22.60	51.83	64.31
		plT5-large[Chrabrowa et al., 2022]	**66.22**	**81.39**	**47.78**	**62.88**	**75.30**

Pytania odpowiadalne

Tabela: Evaluation results on PoQuAD answerable questions only

QA paradigm	Model	HasAns EM	HasAns F1
extractive	mBERT-base	56.76	75.44
	XLM-R-base	60.21	77.59
	XLM-R-large	64.71	82.55
	HerBERT-base	64.21	81.44
	HerBERT-large	**66.43**	**84.38**
abstractive	mT5-base	53.40	70.21
	BART-base	53.68	71.41
	plT5-base	58.30	73.53
	plT5-large	**67.91**	**83.40**

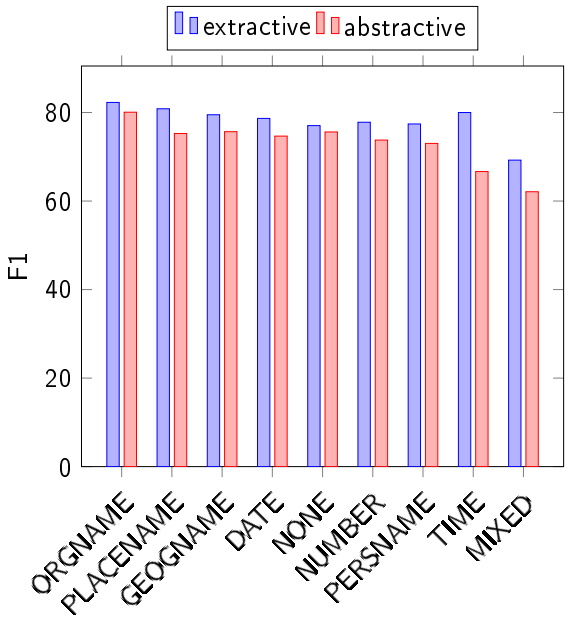
In cases where models failed to recognize the question was unanswerable, the model outputs matched the plausible answers marked by annotators with **F1** of 76.69%.

Pytania nieodpowiadalne

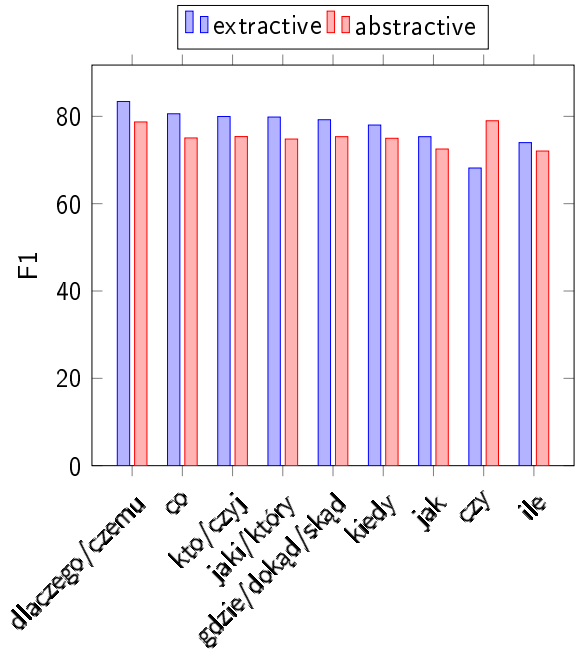
Tabela: Results for answerability classification

Architecture	Train data	Binary F1
TFIDF + Logistic Regression	question	17.95
	question + context	1.07
HerBERT-large binary classifier	question	38.58
	question + context	61.70
HerBERT-large span extractor	original data	65.90

Statystyki odpowiedzi

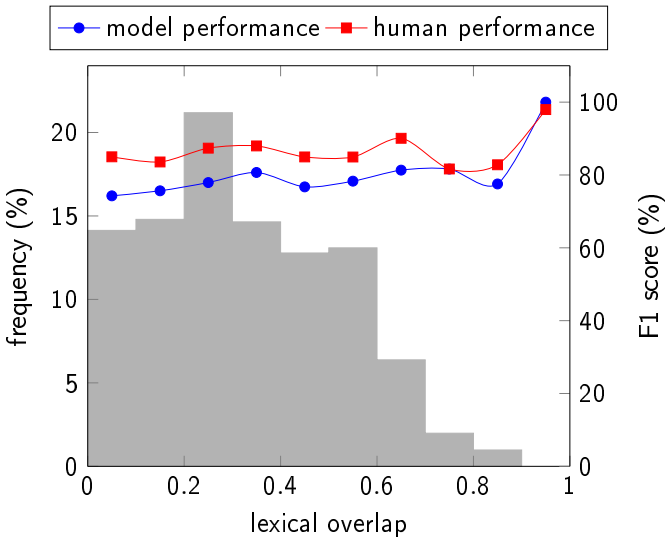


Rysunek: Prediction F1 score by answer entity type

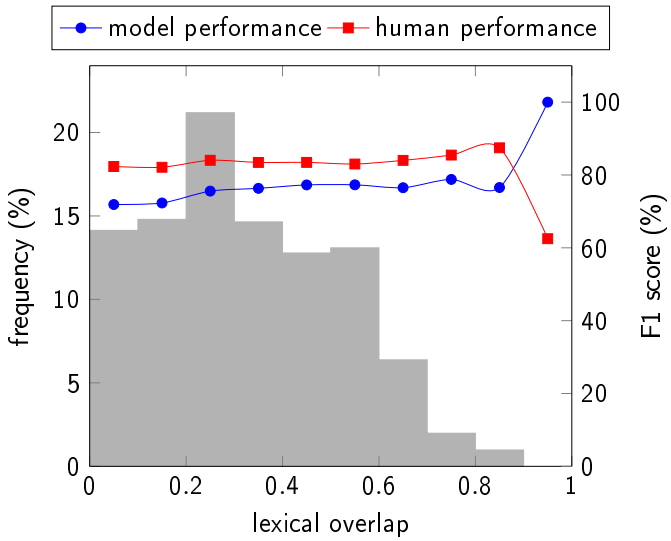


Rysunek: Prediction F1 score by question word

Pokrycie leksykalne



(a) extractive task



(b) abstractive task

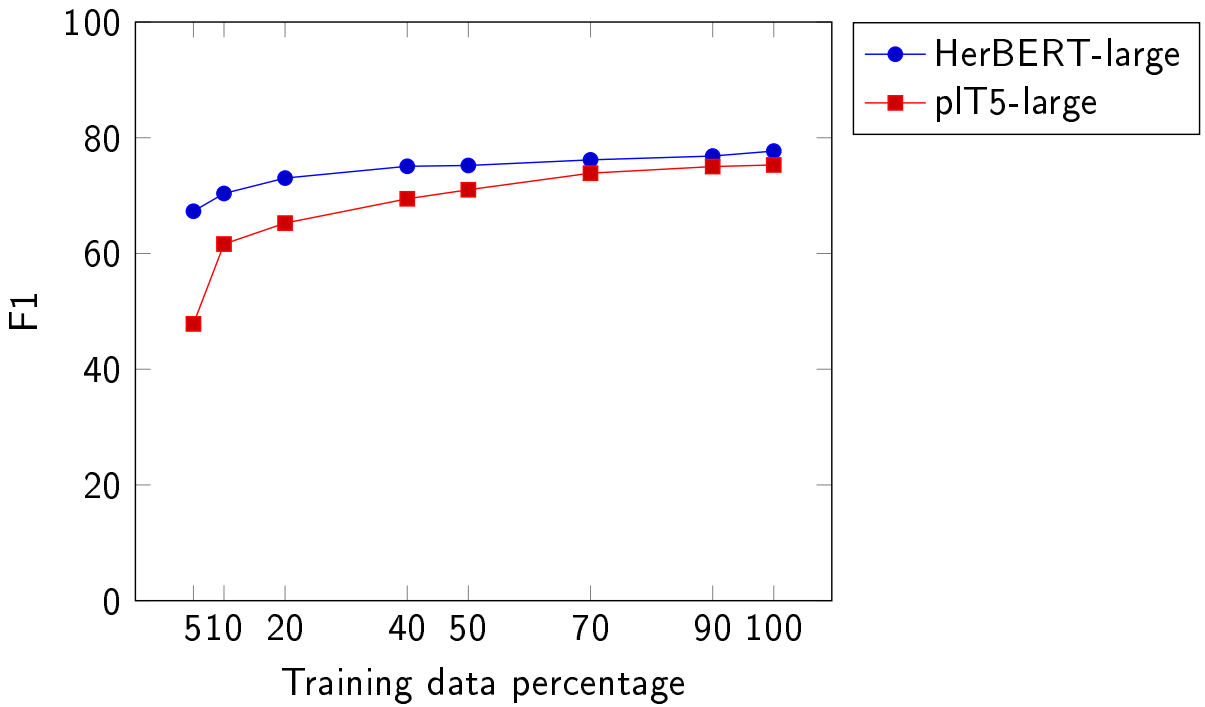
Rysunek: Comparison of human and model performance stratified by lexical overlap (by lemmas) and the distribution of overlap scores. Model performance was assessed on the test subset, while human performance on questions sampled from the dev subset (see **??**). The histogram of lexical overlap was plotted for the full dataset.

Unikalne słowa

Tabela: The effect of unique shared words on model performance (answerable questions only)

Unique word present in the question and the answer sentence	Freq. (%)	F1	
		Extractive	Abstractive
Yes	72.76	82.55	82.93
No	27.24	75.27	77.25

Testy ablacyjne



Rysunek: Results of ablation on train dataset size

Analiza adwersarialna

1. InsertQuestion: randomly inserting the question into the context
2. InsertRandomSent: randomly inserting one of the sentences from the summary into the context
3. OnlyAnswerSents: deleting all sentences from the context except for the ones overlapping with the answer span
4. QuestionWordsOnly: deleting all the words from the question except for the interrogative word
5. ShuffleContext: randomly rearranging the context by sentence
6. ShuffleQuestion: randomly rearranging the question by word

Transformacja możemy nazwać **inwariantną** wtw. gdy odpowiedź na pytanie nie ulega zmianie.

Tabela: Results on adversarially transformed test sets (answerable questions only)

		HerBERT-large		plT5-large	
Manipulation	Invar.	F1	% of F1 lost	F1	% of F1 lost
OnlyAnswerSents	–	80.60	-0.05	81.17	0.27
QuestionWordsOnly	–	2.46	96.95	18.04	77.84
ShuffleContext	–	76.23	5.37	78.84	3.13
ShuffleQuestion	–	66.20	17.83	62.91	22.71
InsertQuestion	+	41.87	48.03	63.14	22.42
InsertRandomSent	+	79.38	1.47	80.91	0.59
baseline		80.56	-	81.39	-

Dziękuję za uwagę!

This work was supported by the European Regional Development Fund as a part of the 2014-2020 Smart Growth Operational Programme: (1) Intelligent travel search system based on natural language understanding algorithms, project no. POIR.01.01.01-00-0798/19; (2) CLARIN — Common Language Resources and Technology Infrastructure, project no. POIR.04.02.00-00C002/19.


Ayoubi, Sajjad & Davoodeh, M. Y. (2021).
Persianqa: a dataset for persian question answering.
https://github.com/SajjjadAyobi/PersianQA.


Borzymowski, H. (2020).
Polish qa model.
model trained on HuggingFace,
https://huggingface.co/henryk/bert-base-multilingual-cased-finetuned-polish-squad2.


Bruyn, M. D., Lotfi, E., Buhmann, J., and Daelemans, W. (2021).
Mfaq: a multilingual faq dataset.


Chrabrowa, A., Dragan, Ł., Grzegorczyk, K., Kajtoch, D., Koszowski, M., Mroczkowski, R., and Rybak, P. (2022).
Evaluation of transfer learning for polish with a text-to-text model.
arXiv preprint arXiv:2205.08808.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., and Stoyanov, V. (2019).
Unsupervised cross-lingual representation learning at scale.
CoRR.

Dadas, S.
Polish BART.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018).
BERT: pre-training of deep bidirectional transformers for language understanding.
CoRR.

d'Hoffschmidt, M., Belblidia, W., Brendlé, T., Heinrich, Q., and Vidal, M. (2020).
Fquad: French question answering dataset.

Efimov, P., Chertok, A., Boytsov, L., and Braslavski, P. (2020).

SberQuAD – russian reading comprehension dataset: Description and analysis.

In *Lecture Notes in Computer Science*, pages 3–15. Springer International Publishing.


Heinrich, Q., Viaud, G., and Belblidia, W. (2022).
FQuAD2.0: French question answering and learning when you don't know.
In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2205–2214, Marseille, France. European Language Resources Association.


Hládek, D., Staš, J., Juhár, J., and Koctúr, T. (2023).
Slovak dataset for multilingual question answering.
IEEE Access, 11:32869–32881.


Lim, S., Kim, M., and Lee, J. (2019).
Korquad1.0: Korean qa dataset for machine reading comprehension.


Liu, J., Lin, Y., Liu, Z., and Sun, M. (2019).
XQA: A cross-lingual open-domain question answering dataset.

In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2358–2368, Florence, Italy. Association for Computational Linguistics.


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


Marcińczuk, M., Radziszewski, A., Piasecki, M., Piasecki, D., and Ptak, M. (2013).
Evaluation of baseline information retrieval for Polish open-domain question answering system.
In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 428–435, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.


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. Association for Computational Linguistics.


Möller, T., Risch, J., and Pietsch, M. (2021).
Germanquad and germandpr: Improving non-english question answering and passage retrieval.


Nguyen, K., Nguyen, V., Nguyen, A., and Nguyen, N. (2020).
A Vietnamese dataset for evaluating machine reading comprehension.
In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2595–2605, Barcelona, Spain (Online). International Committee on Computational Linguistics.


Rajpurkar, P., Jia, R., and Liang, P. (2018).
Know what you don't know: Unanswerable questions for squad.


Rajpurkar, P., Zhang, J., Lopyrev, K., and Liang, P. (2016).
Squad: 100,000+ questions for machine comprehension of text.

In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Rybak, P. (2023).
MAUPQA: Massive automatically-created Polish question answering dataset.
In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 11–16, Dubrovnik, Croatia. Association for Computational Linguistics.

Rybak, P., Przybyła, P., and Ogrodniczuk, M. (2022).
Improving question answering performance through manual annotation: Costs, benefits and strategies.

Sabol, R., Medveď, M., and Horák, A. (2019).
Czech question answering with extended sqad v3.0 benchmark dataset.
In Horák, A., Rychlý, P., and Rambousek, A., editors, *Proceedings of the Thirteenth Workshop on Recent Advances*

in Slavonic Natural Languages Processing, RASLAN 2019, pages 99–108, Brno. Tribun EU.


So, B., Byun, K., Kang, K., and Cho, S. (2022).
Jaquad: Japanese question answering dataset for machine reading comprehension.


Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2020).
mT5: A massively multilingual pre-trained text-to-text transformer.
CoRR.