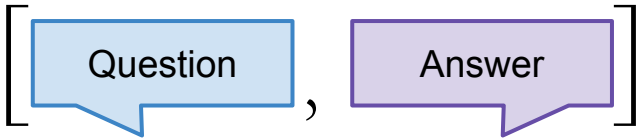
Retrieve and Refine System for Polish Question Answering

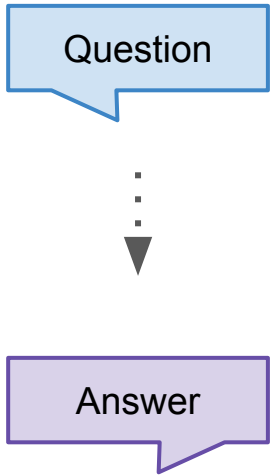
Piotr Rybak, ML Research at Allegro.pl

Datasets

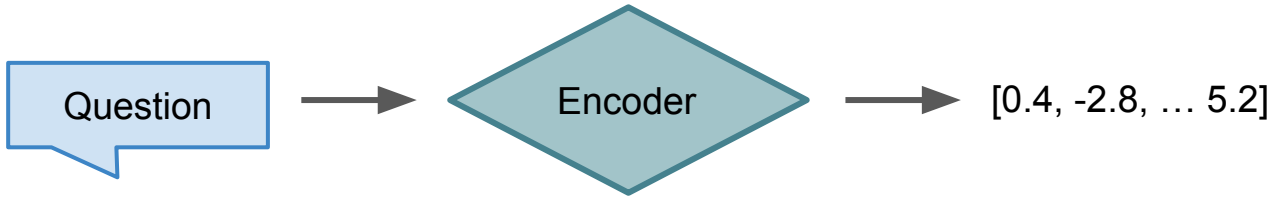
- PolEval
  - Official data provided by task organizers
  - 1000 dev + 2500 test-A
- Jeden z dziesięciu
  - 1004 deduplicated questions
- Multi-lingual Knowledge Questions & Answers
  - Available here: `hf.co/datasets/mkqa`
  - 10k raw questions → 1875 after filtration



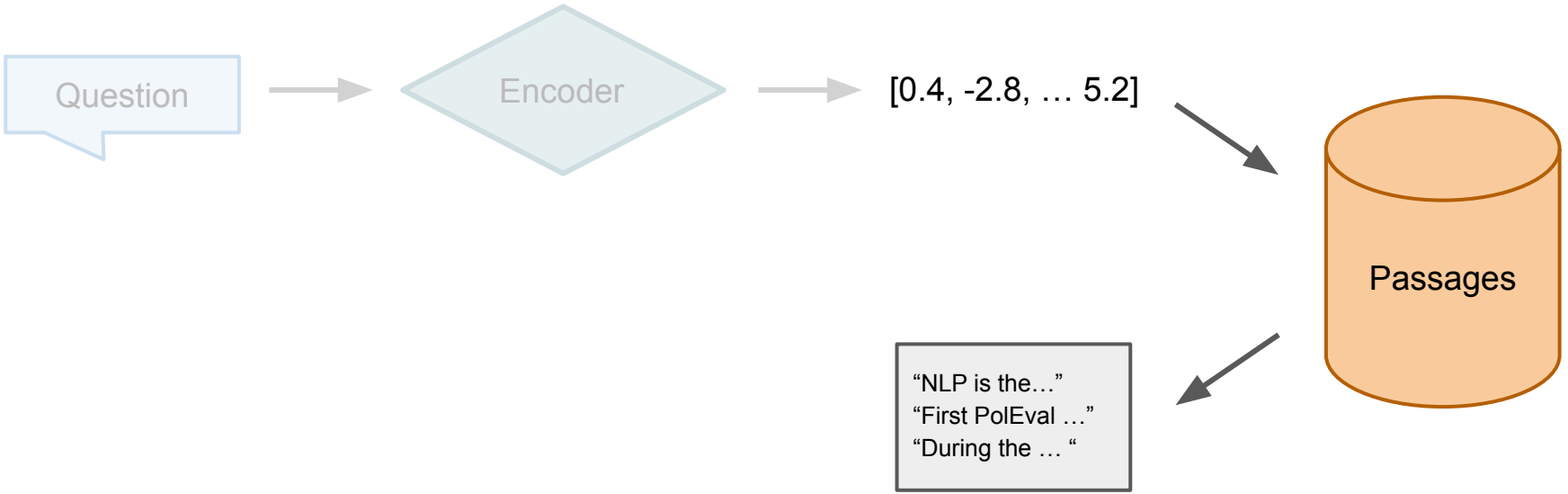
Architecture



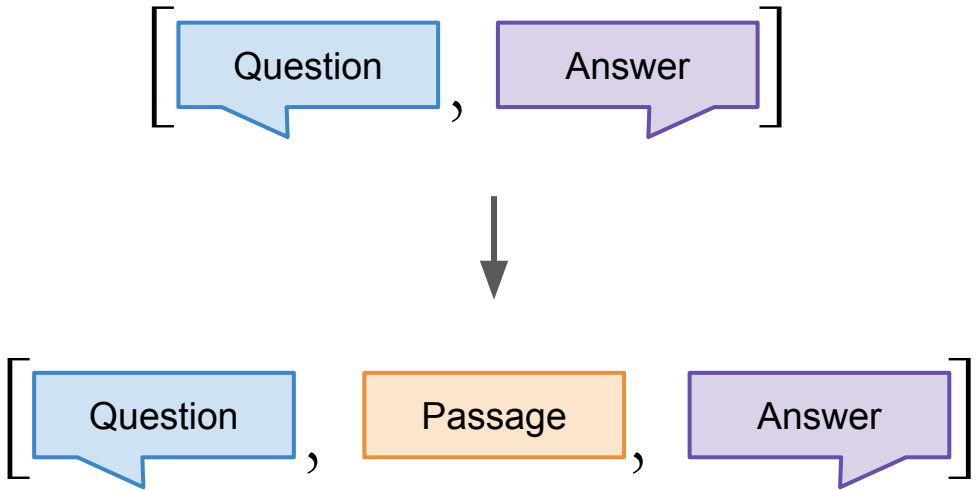
Architecture



Architecture



Question-passage Dataset



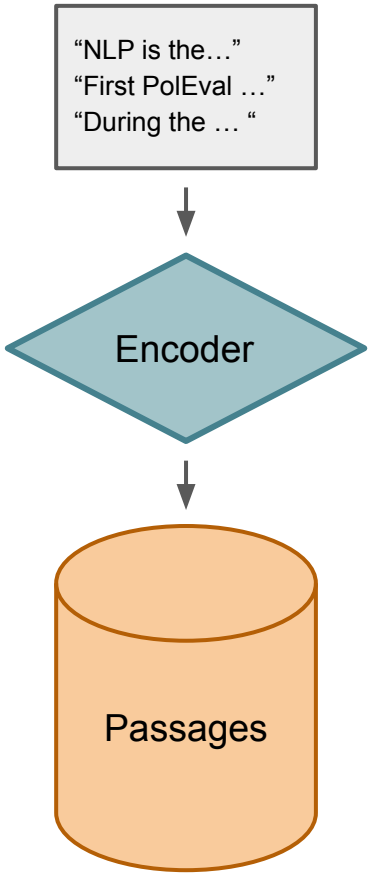
Question-passage Dataset

- Wikipedia
  - Split into paragraphs
  - Concatenated title + paragraph
- Wiktionary
  - Proverbs
  - Idioms

“NLP is the…”
“First PolEval …”
“During the … “

Question-passage Dataset

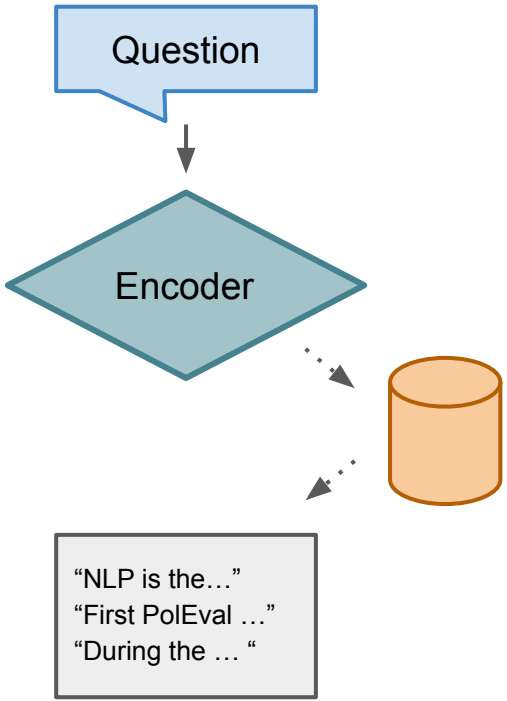
- Encoders
  - Bag-of-Words
  - Universal Sentence Encoder
  - HerBERT trained on “Czy wiesz?”
  - HerBERT trained on part of data
  - HerBERT trained using answers
  - ...



Question-passage Dataset

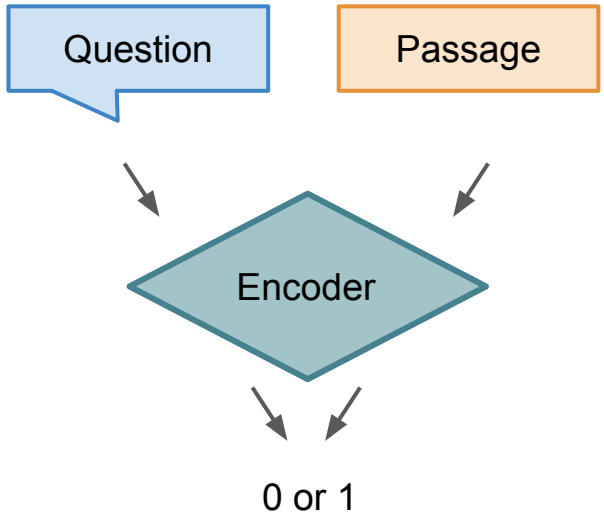
- Manually annotate candidates
  - 10k annotations
  - 2215 positives, 1347 unique questions
  - Available here:

    `hf.co/datasets/allegro/polish-question-passage-pairs`

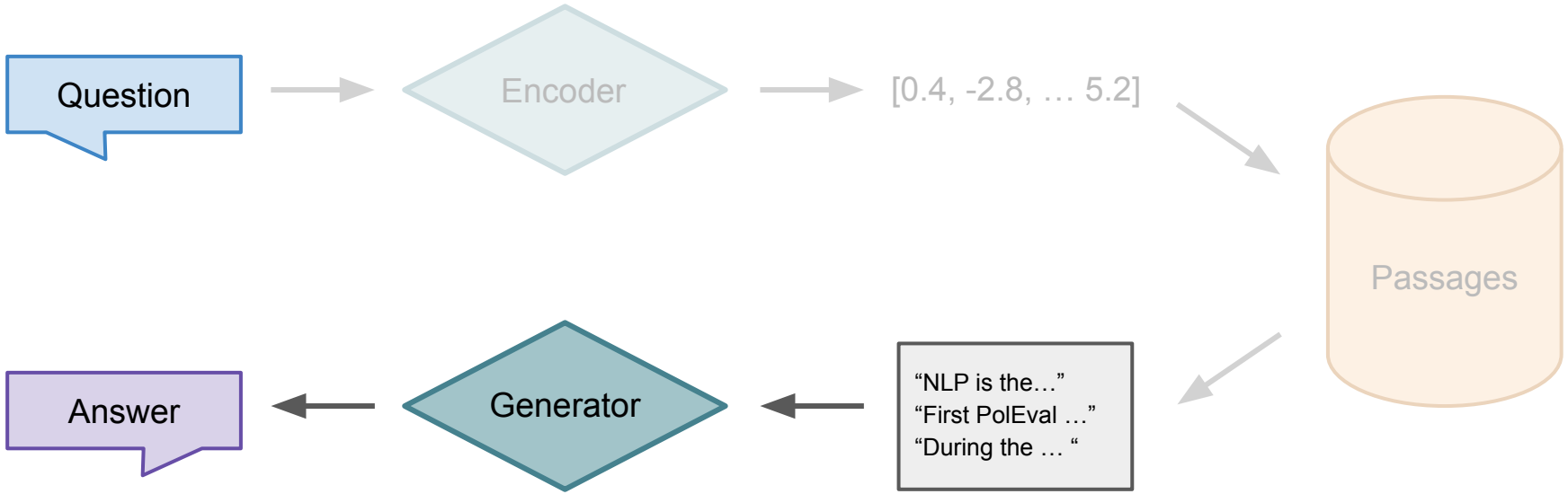


Neural Retriever

- Architecture
  - HerBERT Large
- Dataset
  - 10k annotated pairs
  - “Czy wiesz?” dataset
- Objective
  - Contrastive loss between question & passage

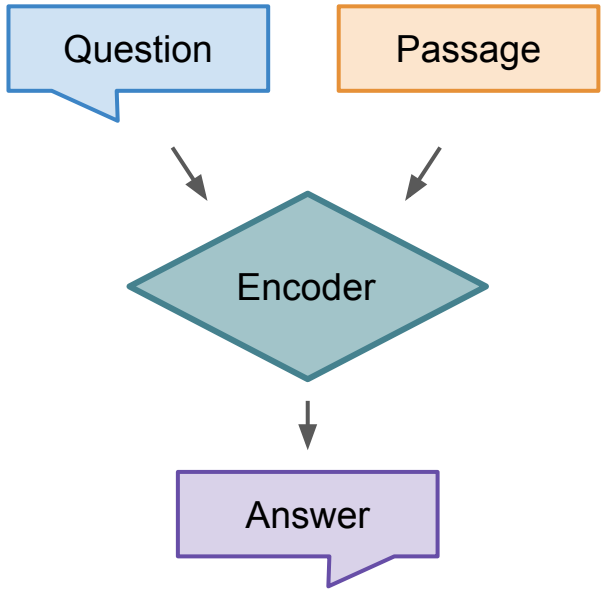


Architecture



Answer Generator

- Architecture
  - plT5 Base
  - `hf.co/allegro/plt5-base`
- First Phase
  - 1347 positives
- Second Phase
  - All question & answer pairs
  - Top 10 predicted passages



Dataset slice	Number of questions	Accuracy
Number	268	50.37
Century	54	85.19
Year	49	61.22
Other	165	35.76
Binary	157	66.88
Yes	97	98.97
No	60	15.00
Definition	417	41.97
Person	304	54.18
Choice	226	58.41
Capital	21	95.24
Proverb	10	80.00
Other	1124	47.06
Total	2500	50.96

Thank you!

piotr.cezary.rybak@gmail.com