Search augmented question answering system using multilangual transformer model

Mateusz Piotrowski

Solution

Two step approach

1. Retrieval of context passages from Wikipedia using Elasticsearch
2. Answer inference with mT5 model (Xue et al. 2021) using question with context as an input

Passage retrieval

Polish Wikipedia corpus was used as a knowledge base. The relevant passages were retrieved using Elasticsearch – search engine based on the Apache Lucene library.

Passage retrieval

Polish Wikipedia corpus was used as a knowledge base. The relevant passages were retrieved using Elasticsearch – search engine based on the Apache Lucene library.

1. Split articles from Wikipedia dump into paragraphs

Passage retrieval

Polish Wikipedia corpus was used as a knowledge base. The relevant passages were retrieved using Elasticsearch – search engine based on the Apache Lucene library.

1. Split articles from Wikipedia dump into paragraphs
2. Perform lemmatization using Polish spaCy model

Passage retrieval

Polish Wikipedia corpus was used as a knowledge base. The relevant passages were retrieved using Elasticsearch – search engine based on the Apache Lucene library.

1. Split articles from Wikipedia dump into paragraphs
2. Perform lemmatization using Polish spaCy model
3. Index documents (both forms) in Elasticseach

Passage retrieval

Polish Wikipedia corpus was used as a knowledge base. The relevant passages were retrieved using Elasticsearch – search engine based on the Apache Lucene library.

1. Split articles from Wikipedia dump into paragraphs
2. Perform lemmatization using Polish spaCy model
3. Index documents (both forms) in Elasticseach
4. Retrieve passages using lemmatized question as a query

Passage retrieval

Polish Wikipedia corpus was used as a knowledge base. The relevant passages were retrieved using Elasticsearch – search engine based on the Apache Lucene library.

1. Split articles from Wikipedia dump into paragraphs
2. Perform lemmatization using Polish spaCy model
3. Index documents (both forms) in Elasticseach
4. Retrieve passages using lemmatized question as a query
5. Concatenate question and top N results up to a token limit to obtain input prompt

Example prompt

Czyją żoną w „Chłopach” Reymonta była Jagna?

context

Jagna Jagienka ze Zgorzelic - bohaterka powieści Krzyżacy H. Sienkiewicza Jagna Borynowa, z d. Pacześ - bohaterka Chłopów Wł. St. Reymonta Jagienka Oracabessa - kilkuletnia czarnoskóra bohaterka powieści Język Trolli Małgorzaty Musierowicz Jagna Obłoczek - postać z filmu ”Zwierzogród” Magdalena Koleśnik W 2014 roku otrzymała nagrodę za rolę Marii Łukianowny w spektaklu „Samobójca” na 32. Festiwalu Szkół Teatralnych w Łodzi, w 2015 roku Nagrodę im. Andrzeja Nardellego za najlepszy debiut aktorski; za rolę w spektaklu ”Dybuk”, zaś w 2017 roku nagrodę dla Młodego Twórcy za rolę Jagny z ”Chłopów” według Władysława Reymonta w reżyserii Krzysztofa Garbaczewskiego z Teatru Powszechnego w Warszawie ”ze względu na wykreowanie niebywale przejmującej roli Jagny, która dzięki imponującej precyzji warsztatu aktorskiego, przełamuje klasyczne uwizerunkowienie postaci” na 4. Festiwalu Nowego Teatru w Rzeszowie. Tomasz Jodełka-Burzecki Twórczość Jan Kasprowicz. Zarys biografii Nad tekstami ”Chłopów” Reymonta Reymont przy biurku. Z zagadnień warsztatu pisarskiego Władysław Reymont W to tło wpleciona jest główna oś akcji powieści, którą jest romans młodej, urodziwej i namiętnej **Jagny - żony bogatego gospodarza Macieja Boryny**, z jego synem Antkiem oraz postawy społeczności chłopskiej wobec tego zdarzenia. Po śmierci Macieja Boryny (który stał na czele buntu chłopów wobec dworu), Jagna zostaje napiętnowana i wypędzona przez społeczność wiejską i ulega obłędowi, zaś

Answer inference

The pre-trained mT5 model was used to infer the answer from provided context.

Answer inference

The pre-trained mT5 model was used to infer the answer from provided context.

- Encoder-decoder transformer model is capable of generating arbitrary text

Answer inference

The pre-trained mT5 model was used to infer the answer from provided context.

- Encoder-decoder transformer model is capable of generating arbitrary text
- Multilingual model can be fine-tuned using a wider array of datasets

Answer inference

The pre-trained mT5 model was used to infer the answer from provided context.

- Encoder-decoder transformer model is capable of generating arbitrary text
- Multilingual model can be fine-tuned using a wider array of datasets
- Pre-trained language models have access to knowledge encoded in neural network weights (Roberts, Raffel, and Shazeer 2020)

Additional training data

Combination of questions from *dev-0* and *test-A* sets were used to create base dataset consiting of 3500 examples. Extending the training data with 87K+ examples from English TriviaQA dataset (Joshi et al. 2017) had more significant impact on smaller model

model	PL only	combined
base	47.64	52.12
large	58.12	59.20
XXL	70.76	71.68

Closed book QA

Results on test-B set in closed-book setting.

model	no context
base	17.00
large	22.76
XXL	36.20

Closed-book QA

Anion posiada przewagę elektronów czy protonów?

base: protonów

large: elektronów

XXL: protonów

Closed-book QA

W którym roku miała miejsce rewolucja październikowa w Rosji?

base: w 1989

large: w 1917

XXL: w 1917

Closed-book QA

Jak miał na imię Gomułka, pierwszy sekretarz PZPR?

base: Zbigniew

large: Sławomir

XXL: Władysław

Conclusions

- Model size is a significant factor in overall performance
- Passage retrieval recall might be improved by using search methods based on dense document representations instead of sparse TF-IDF (Guu et al. 2020, Karpukhin et al. 2020)
- Modifying pre-training objective to focus on real-world knowledge might improve accuracy in QA tasks (Roberts, Raffel, and Shazeer 2020)

Acknowledgments

Research supported with Cloud TPUs from Google's TPU Research Cloud (TRC)

Thank you for your attention!

References

[Guu+20] Kelvin Guu et al. *REALM: Retrieval-Augmented Language Model Pre-Training*. 2020. arXiv: 2002.08909 [cs.CL].

[Jos+17] Mandar Joshi et al. *TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension*. 2017. arXiv: 1705.03551 [cs.CL].

[Kar+20] Vladimir Karpukhin et al. *Dense Passage Retrieval for Open-Domain Question Answering*. 2020. arXiv: 2004.04906 [cs.CL].

[RRS20] Adam Roberts, Colin Raffel, and Noam Shazeer. “How Much Knowledge Can You Pack Into the Parameters of a Language Model?” In: *Empirical Methods in Natural Language Processing (EMNLP)*. 2020.

[Xue+21] Linting Xue et al. “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*. Online: Association for Computational Linguistics, June 2021, pp. 483–498. DOI: 10.18653/v1/2021.naacl-main.41. URL: https://aclanthology.org/2021.naacl-main.41.