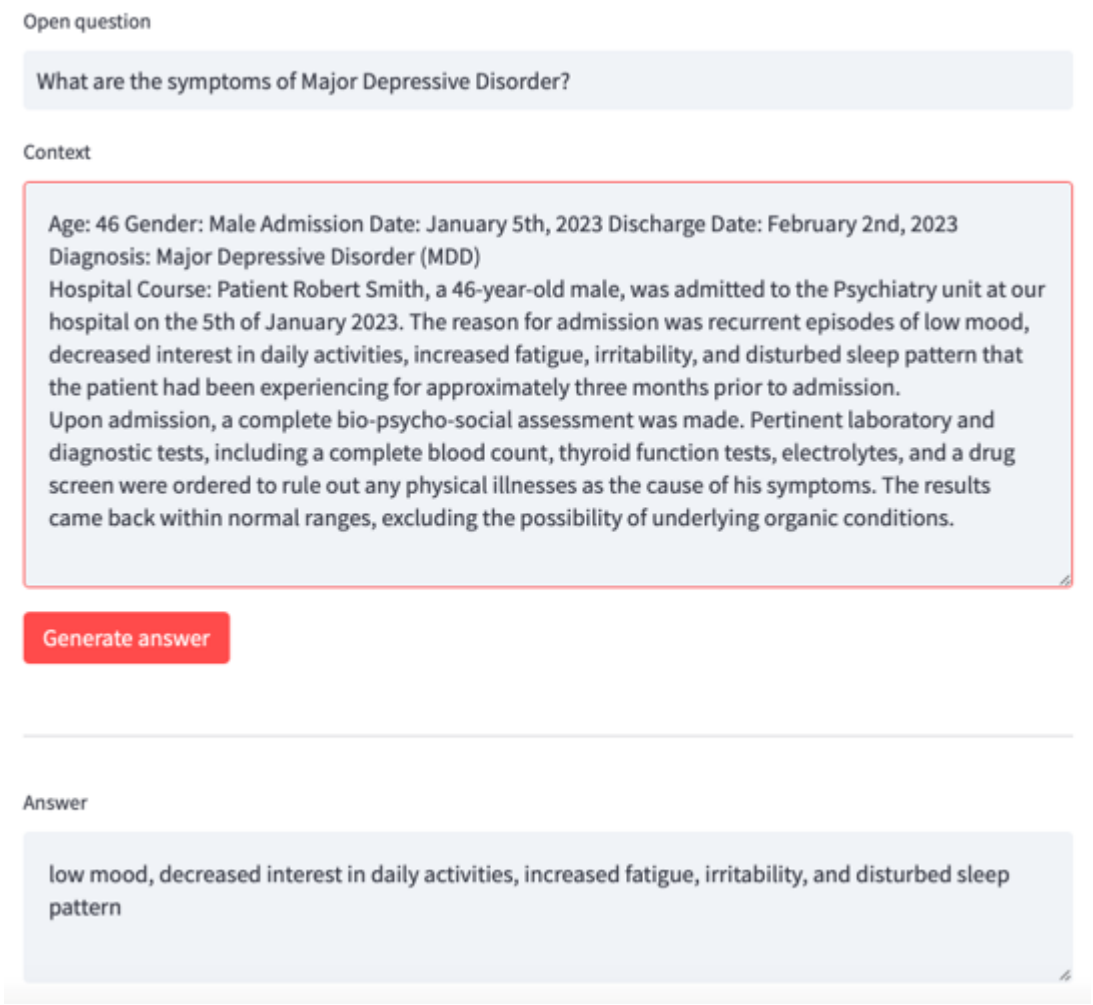
ontotext

This is part of Ontotext's AI-in-Action initiative aimed at enabling data scientists and engineers to benefit from the AI capabilities of our products.

Have you ever tried to find out what your medical records say after a doctor's visit? Probably most of you have tried, but how many could figure out all the details of the described information – symptoms, laboratory tests, accompanying diseases, therapy, and so on? Even for a human, this task is difficult to solve.

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com



Challenges

Medical multilingual question answering (QA) presents several challenges stemming from the diverse nature of medical terminologies and linguistic variations. One crucial issue is the lack of standardized medical terminology translation, which can lead to incorrect or ambiguous interpretations of the terms.

Moreover, the complexity of medical language and concepts requires QA models to have a deep understanding of context. This makes it challenging to develop robust systems that can accurately interpret queries and provide relevant answers across languages. Furthermore, as

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

ontotext

the clinical data is highly sensitive, there are no open-access models or datasets available to solve the task, especially in a multilingual setting.

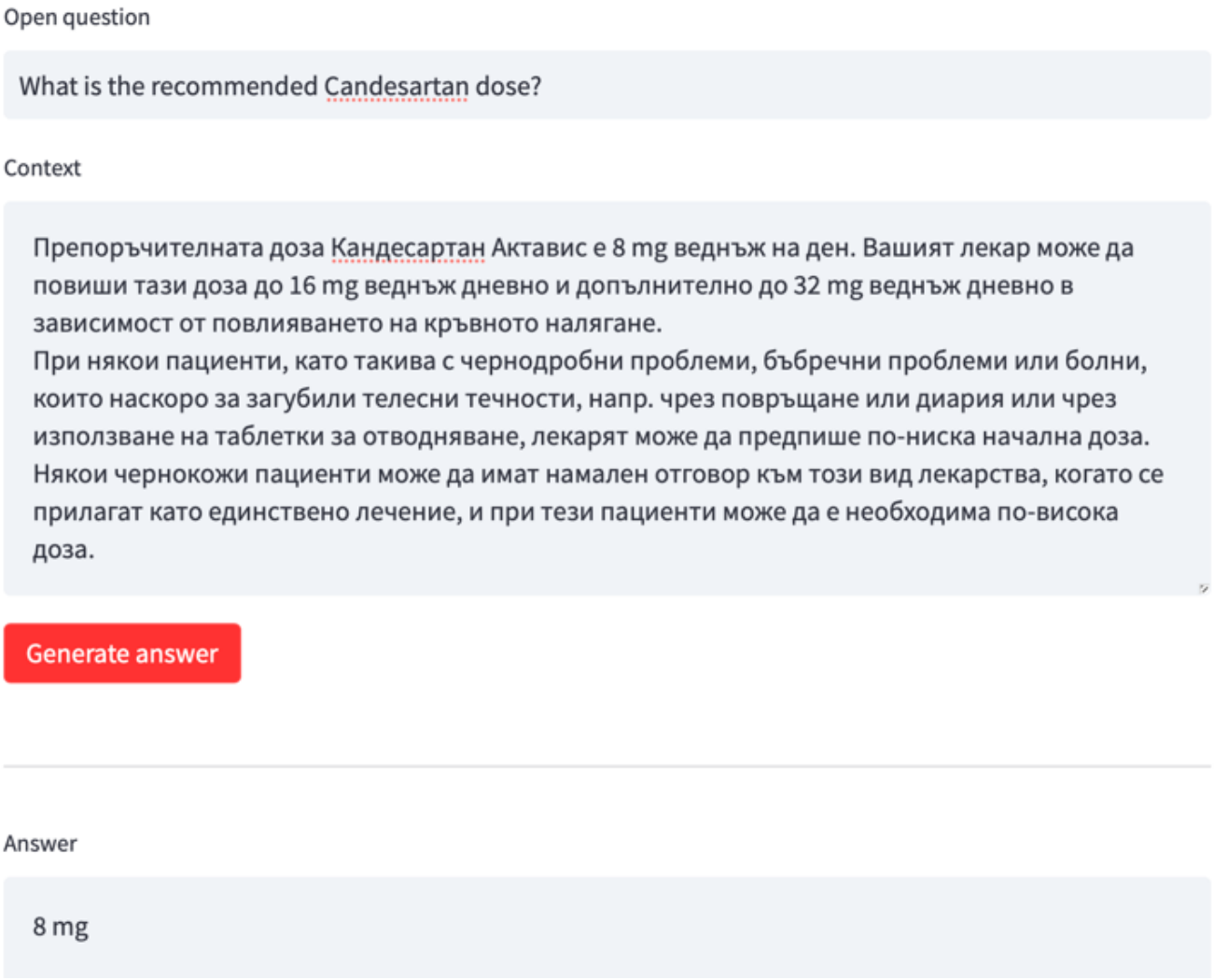
Multilinguality

In applying extractive QA, multilinguality and language-agnostic knowledge transfer is of great benefit. Imagine you can ask questions in English, Spanish, or any language you choose, and the system understands you and provides the answer from the document where you search in any language. The specifics of training such models don't require data in all the languages that you might target but a set of diverse examples.

For instance, in one of our research projects, RES-Q PLUS, we are facing very low resource multilinguality in applying information extraction methods for medical documents in Bulgarian, Czech, English, Greek, Polish, Romanian, and Spanish. There isn't enough annotated data that would allow us to train named entity recognition (NER). But with the QA approach, we can easily extract more than a hundred concepts of interest and even model the relationships between them.

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

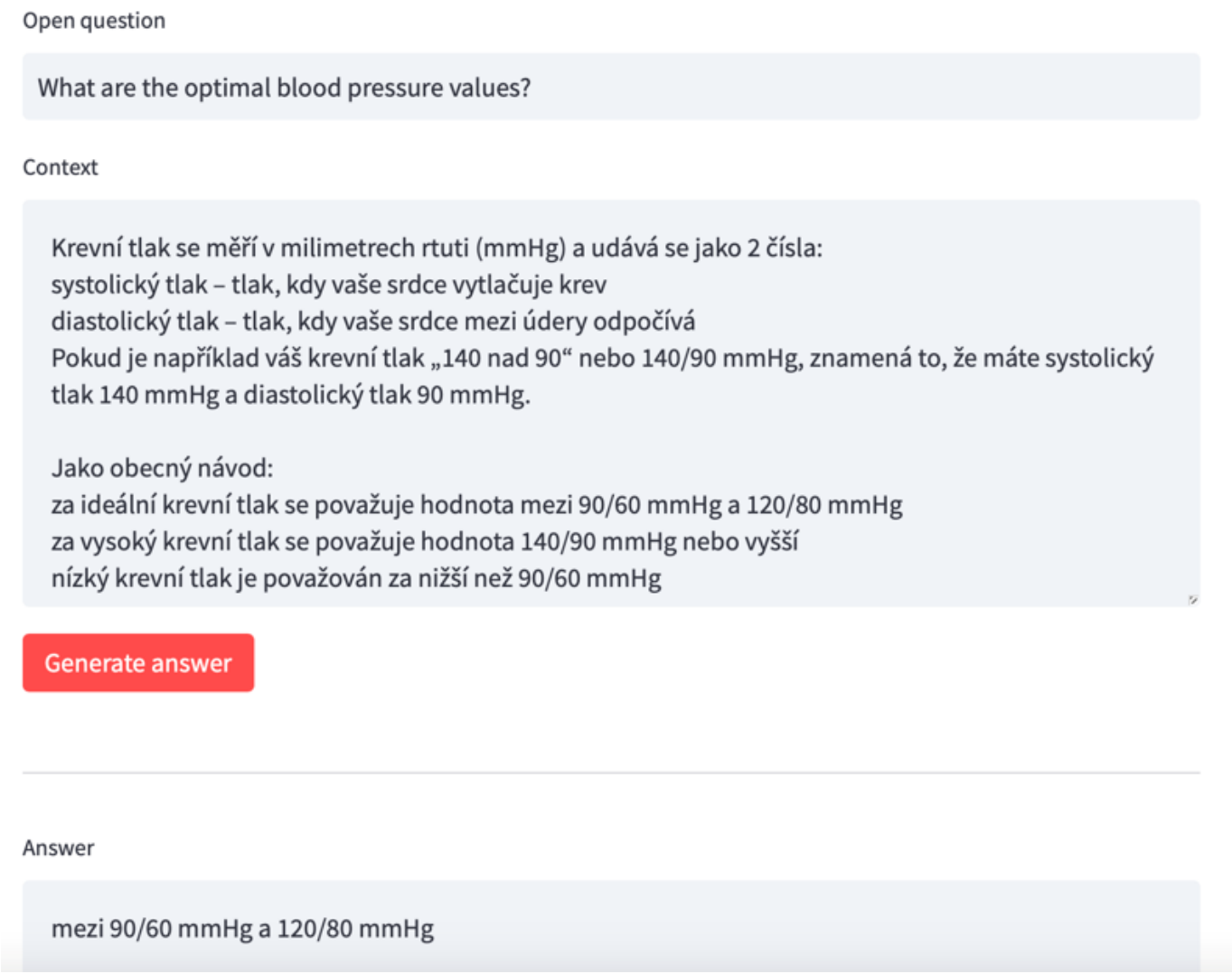
ontotext



An example of extraction methods for medical documents in Bulgarian

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

ontotext



An example of extraction methods for medical documents in Czech

Under the hood

The magic happens with the help of a transformer-based model and several carefully curated training steps.

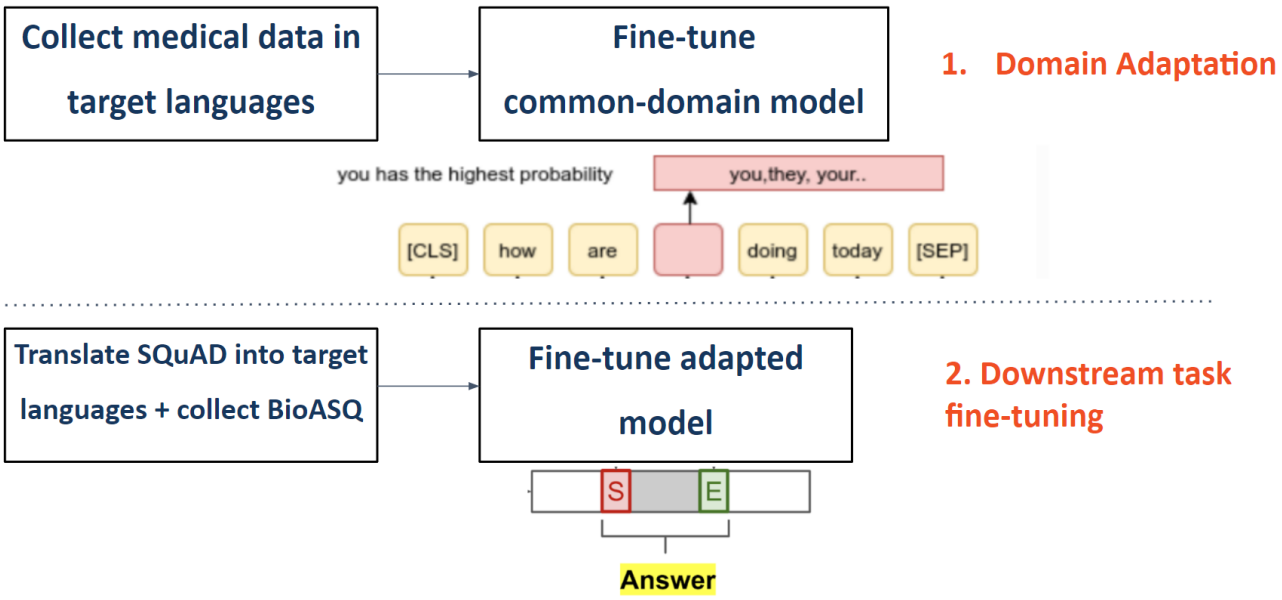
Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

ontotext

Domain Adaptation

To adapt the model for specific needs, we should first train it on the set of raw in-domain data with masked language modeling objective and then on the dataset that would represent the target extraction task. Masked language modeling in domain adaptation of transformers involves fine-tuning the pre-trained transformer model by masking certain tokens in the input text and predicting them based on the surrounding context. This process helps the transformer model effectively learn domain-specific patterns and nuances. It enhances its performance when applied to tasks within the target domain while also leveraging the knowledge obtained during the pre-training on common-domain data.

For this task, we gathered open-access data related to the medical domain in target languages including drug descriptions, medical vocabularies and thesauri, Wikipedia articles, machine-translated medical data, and so on. The XLM-RoBERTa model was trained on this corpus as a simple yet efficient multilingual backbone used for many medical natural language processing (NLP) tasks.



Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

ontotext

QA fine-tuning

The next step in the training pipeline is task-specific fine-tuning. Originally, extractive QA is set up as a token classification problem where for any token in the text we are learning whether it is the start of the answer, the end of the answer, or neither. Such a setting gives us a set of advantages over NER. As it generally solves a simpler optimization problem (the number of token classes in NER is at least twice as large as the number of entity types), it requires less data to train. In addition, the task setup is question-agnostic, therefore transfer-learning in this case is very efficient.

The most famous extractive QA dataset that is widely used to benchmark language models is the Stanford Question Answering Dataset (SQuAD). The problem is that this dataset doesn't really fit our multilingual setting as it was created exclusively for English. For the purposes of RES-Q PLUS, we found or created SQuAD versions for all the target languages.

In the biomedical domain, one of the datasets created similarly to SQuAD is BioASQ. We used the challenge 7b version transformed into SQuAD format and combined it with the 2023 -11b questions preprocessed by us for the challenge. We also attempted to apply machine translation for this data. However, the answers were significantly misaligned and the algorithm didn't produce good results. Therefore we are using the English-only version of this dataset.

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

ontotext

Main Takeaways

In conclusion, extractive QA proved a promising technology. It brings information extraction to the next level by allowing more flexibility in the search for information. Its ability to analyze not only the closed set of entities but also find the answer to questions makes it more powerful for connecting various pieces of information.

The adaptation pipeline can be replicated for any set of languages and domain-specific data. This makes it a good fit for information extraction projects where the reliability of the answers is key. For examples of other domains and data, please check our fundamental: What is Extractive Question Answering?.

Some next steps can be applied for transforming extracted data into a structured format or a knowledge graph. These may include post-processing steps such as entity linking to medical ontologies, using some of the rich resources available in LinkedLifeData Inventory.

Another possible utilization of the extractive QA model is in an NLP pipeline integrated into Ontotext Metadata Studio for automatic annotation of clinical texts.

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

ontotext

Do you want to see what Ontotext's multilingual extractive question answering is capable of?

Book a Discovery Call

Ontotext's work and experiments with multilingual extractive question answering models have been carried out as part of the RES-Q+ project.

Copyright: Ontotext New York | Sofia | Basel
Reach out to us: info@ontotext.com

Multilingual Question Answering in Medicine based on XLM-RoBERTa

Author: Anna Aksenova

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No:101057603. Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union. Neither the European Union nor the granting authority can be held responsible for them.

Copyright: Ontotext New York | Sofia | Basel

Reach out to us: info@ontotext.com