Creation of a Training Dataset for Question-Answering Models in Serbian

Scientific paper DOI: 10.18485/judig.2025.1.ch14

Ranka Stanković, *0000-0001-5123-6273*
Jovana Rađenović, *0000-0002-2707-3870*
Maja Ristić, *0000-0002-3539-018X*
Dragan Stankov[1] *0000-0001-9394-9247*

Abstract

The paper presents an overview of the different formats and domains of both multilingual and monolingual resources for the task of Question Answering (QA), with a particular focus on Serbian. Within the TESLA (Text Embeddings – Serbian Language Applications) project, a dataset consisting of context, questions, and answers is being developed with sources taken from various domains. The structure for the QA dataset, composed of four smaller subsets, and the method for its development will be outlined in this paper. The development of the first subset was based on an adaptation of a subset of the Stanford set SQuAD. The second subset *TESLA-Sveznanje-QA* was generated from the Sveznanje Encyclopedia, using LLMs, while the third part *TESLA-domain-QA* is produced from textbooks, also using LLMs. The fourth subset *TESLA-Wikidata-QA* contains automatically generated contexts based on the content of the Wikidata knowledge base and LLMs. The research conclusion indicates the importance and potential of applying this dataset in various fields, including educational technologies, digital assistants, and information retrieval systems. The presented results contribute to the improvement of language technologies for Serbian, and we hope that they will encourage further research and development in this area.

[1] University of Belgrade, Faculty of Mining and Geology
E-mail: {ranka.stankovic|jovana.radjenovic|maja.ristic|dragan.stankov}@rgf.bg.ac.rs

Keywords: artificial intelligence, natural language processing, language resources, annotated sets, information extraction, question answering

1. Introduction

The development and application of artificial intelligence in language technologies have advanced significantly in recent years, especially in the domain of the task of answering questions (Question Answering - QA). While existing resources for QA tasks have been developed for major world languages, the Serbian language has been relatively neglected in this area. This work represents an initiative to create an extensive and diverse set of data for training models to answer questions in Serbian, thus contributing to the improvement of language technologies for Serbian.

In addition to the numerous studies on language models in the last few years, much work has also been done on the reference datasets needed to track the modelling progress. A lot has been done when it comes to answering questions and understanding what is read, although mostly when it comes to big languages (Rogers et al. 2023). We will provide an overview of the various formats and domains of available multilingual and monolingual resources, with special reference to Serbian (Cenić & Stojković 2023; Cvetanović & Tadić 2023, 2024).

The construction of high-quality question-answer (QA) datasets is essential for the development and evaluation of machine learning models, particularly for natural language understanding (NLU) and retrieval-based tasks. Traditional QA dataset creation methods rely on manual curation, rule-based extraction, or semi-automated approaches using structured knowledge bases. However, recent advancements in Large Language Models (LLMs) provide new opportunities for generating diverse and contextually rich QA pairs from textual corpora. Chen et al. (2019) show that while existing n-gram based metrics (BLEU, ROUGE, METEOR, F1) are somewhat suitable for current QA datasets, they significantly limit the development of more complex, free-form QA tasks and do not always correlate well with human response, suggesting a need for new metrics such as adapted BERT-based approaches.

Question Answering (QA) systems enable users to retrieve information from various sources, including both structured and unstructured data in natural language. As a crucial component of Conversational AI, QA has led to the emergence of Conversational Question Answering (CQA), where a system needs to understand the context and engage in multi-turn interactions to address user queries effectively. While most existing research has focused on single-turn QA, multi-turn QA has gained increasing attention

due to the availability of large-scale datasets and advances in pre-trained language models. With a growing number of studies contributing to the field each year, there is a pressing need to organize and synthesize existing research to guide future work. Zaib et al. (2022) provided a comprehensive review of state-of-the-art CQA research, highlighting a clear shift from single-turn to multi-turn QA, which enhances Conversational AI in multiple ways.

Knowledge Graph Question Answering (KGQA) has seen growing interest, with numerous benchmarking datasets driving advancements in the field. While earlier benchmarks relied on Freebase and DBpedia, research has shifted toward Wikidata due to its superior structural validity. In response, a new multilingual, complex KGQA benchmark was introduced as part of the QALD-10 series, transitioning from DBpedia to Wikidata. This adaptation required addressing challenges such as data complexity, cross-language mapping, and property ranking. Usbeck et al. (2024) presented a case study, conducted as a conference challenge, providing insights into the benchmark's creation and its role in advancing KGQA research. Knowledge Base Question Generation (KBQG) involves generating questions from structured database information, typically represented as triples. Hun et al. (2022) work on KBQG using pre-training, a new (triple, question) dataset, and question-type classification, demonstrating improved performance in standard and zero-shot settings. The extended KBQG2 dataset enhances knowledge base coverage and increases output variability, enabling multiple question formulations for the same KB triple.

Korablinov et al. (2020) introduced RuBQ, the first Russian Knowledge Base Question Answering (KBQA) dataset, containing 1,500 Russian questions of varying complexity, their English translations, SPARQL queries to Wikidata, reference answers, and a Wikidata sample with Russian-labeled entities. The dataset was created from online quiz question-answer pairs, which were subject to automatic filtering, crowd-assisted entity linking, automated SPARQL query generation, and in-house verification. RuBQ is a freely available and valuable resource for researchers in Semantic Web, NLP, and Information Retrieval, particularly in multilingual question answering. The dataset generation pipeline proved to be efficient and can be adapted for other data annotation projects.

QA datasets can be broadly categorized based on their structure, data sources, and the nature of the questions they contain. Extractive QA datasets, such as SQuAD[3] and Natural Questions[4], require models to identify answers

[2] The code and dataset are publicly available at https://gitlab.inria.fr/hankelvin/wikidataqg
[3] https://rajpurkar.github.io/SQuAD-explorer/
[4] https://ai.google.com/research/NaturalQuestions

directly within a given passage. In contrast, abstractive QA datasets, like NarrativeQA[5] and certain parts of the MS MARCO dataset[6], demand generative responses that go beyond exact text spans, requiring deeper language comprehension and rephrasing capabilities. Another major category is KBQA, where datasets such as WebQuestions[7] and MetaQA[8] rely on structured knowledge bases like Wikidata to retrieve factual answers.

With the rise of conversational AI, Conversational QA (CQA) datasets have become increasingly significant. Datasets like CoQA[9] and QuAC[10] enable models to engage in multi-turn interactions, maintaining context across multiple exchanges. Meanwhile, open-domain QA datasets, such as TriviaQA[11] and HotpotQA[12], allow models to search for answers across large, unstructured text corpora, often requiring retrieval from multiple sources. Multi-hop QA datasets, including HotpotQA and ComplexWebQuestions[13], add another layer of complexity by requiring reasoning across multiple documents to arrive at a final answer.

Beyond factual retrieval, some QA datasets focus on commonsense reasoning, such as CommonsenseQA[14] and Social IQA[15], which challenge models to infer implicit knowledge beyond what is explicitly stated in text. Similarly, multimodal QA datasets, such as VQA[16] and TextVQA[17], extend the scope of QA to include images and documents, requiring an understanding of both textual and visual elements. Domain-specific datasets, such as biomedical QA datasets like BioASQ[18] and MedQA[19], address specialized fields where expert-level reasoning is required. Additionally, code-related QA datasets, such as CoNaLa[20] and CodeSearchNet[21], are designed for programming-related queries, assisting in automatic code generation and problem-solving.

As QA research progresses, the diversity of available datasets continues to expand, enabling models to tackle increasingly sophisticated

[5] https://github.com/google-deepmind/narrativeqa
[6] https://microsoft.github.io/msmarco/
[7] https://github.com/brmson/dataset-factoid-webquestions
[8] https://github.com/yuyuz/MetaQA?tab=readme-ov-file
[9] https://stanfordnlp.github.io/coqa/
[10] https://quac.ai/
[11] https://nlp.cs.washington.edu/triviaqa/
[12] https://hotpotqa.github.io/
[13] https://www.tau-nlp.sites.tau.ac.il/compwebq
[14] https://www.tau-nlp.sites.tau.ac.il/commonsenseqa
[15] https://maartensap.com/social-iqa/
[16] https://visualqa.org/
[17] http://textvqa.org/dataset/
[18] https://participants-area.bioasq.org/datasets/
[19] https://github.com/jind11/MedQA
[20] https://conala-corpus.github.io/
[21] https://paperswithcode.com/dataset/codesearchnet

questions. From simple fact-based queries to complex multi-step reasoning and conversational interactions, the development of high-quality QA datasets remains essential for enhancing the capabilities of AI-driven question-answering systems across various domains. Hopkins et al. (2019) presented results of the SemEval 2019 task on math question answering, utilizing a dataset derived from Math SAT practice exams with logical form annotations for a subset, which reached an accuracy of 45%, thus significantly outperforming a random guessing baseline.

As part of the TESLA (Text Embeddings - Serbian Language Applications) project, we are working on the preparation of a set of data: context, questions, and answers, collected from different domains. In this paper, we will explain several approaches that we are using for QA dataset development in the TESLA project. The approach for the development of the first dataset, based on the Stanford SQuAD set (Rajpurkar et al., 2018), will be described in Section 2. The TESLA-Sveznanje-QA dataset, derived from encyclopedic knowledge, will be presented in Section 3.

Section 4 is dedicated to TESLA-domain-QA, a domain-specific QA with a dataset containing at least 5000 questions with answers and the relevant context, excerpted from a textbook published by the Faculty of Mining and Geology. The last dataset, TESLA-Wikidata-QA contains automatically generated contexts based on the content of the Wikidata knowledge base and Gemini 4.0. Its development is explained in Section 5, which outlines a method for transforming structured knowledge into a high-quality question-answering system.

2. TESLA-SQuAD-sr Dataset

In the process of constructing the first dataset, we turned to the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2018) as the basic reference. SQuAD consists of textual contexts extracted from Wikipedia articles, with questions and answers manually created by human annotators based on the given context. This dataset is widely used for training and evaluating machine learning models in question-answering tasks, providing a well-structured benchmark with high linguistic quality.

Following this approach, we analyzed the SQuAD-sr dataset (Cvetanović & Tadić, 2024), which was developed by applying a semi-automated translation approach to the original SQuAD dataset. This process involved translating the English content into Serbian while ensuring that the structure of questions, answers, and contextual passages was preserved. The use of semi-automated methods allowed for a faster adaptation of an existing QA dataset into Serbian, while also introducing potential translation

inconsistencies that required manual correction. Table 1 provides the average (avg), maximum (max), and minimum (min) character count for each element in the SQuAD v1.1 and SQuAD-sr datasets. The primary purpose of these statistics is to enable a thorough analysis of text length within both datasets. This, in turn, facilitates a direct comparison of textual characteristics between the English and Serbian versions, offering insights into potential linguistic and/or structural differences.

Table 1. Dataset metrics SQuAD v1.1 and SQuAD-sr (in characters)

	SQuAD v1.1			**SQuAD-sr**		
	avg	**max**	**min**	**avg**	**max**	**min**
context	736	3706	151	715	3520	151
question	60	270	1	57	1063	0
answer	20	239	1	20	570	1

The first part of our dataset, *TESLA-SQuAD-sr*, represents a refined subset of SQuAD-sr, with corrections to improve translation accuracy, linguistic coherence, and domain relevance. The selected subset comprises topics such as information technology, climate change, civil engineering, geology, and similar areas. These topics were chosen to ensure a diverse yet specialized knowledge base that can support QA system development in both general and domain-specific contexts.

The finalized subset will contain approximately 7,000 questions with corresponding answers, offering a structured and high-quality dataset suitable for further model training and evaluation. This refined resource aims to contribute to the development of Serbian-language QA systems by enhancing both linguistic precision and contextual depth in machine-learning applications.

Currently, the dataset contains 5,000 QA pairs across 36 different topics. The original *SQuAD* dataset contains 87,599 QA pairs related to 442 topics, and the *SQuAD-sr* dataset contains 87,175 QA pairs for the same number of topics. Topics from *SQuAD* that weren't of interest, such as Sino-Tibetan relations during the Ming dynasty, The Legend of Zelda: Twilight Princess, Southampton, Catalan language, Estonian language, etc., were not taken into consideration for *TESLA-SQuAD-sr*. The questions in our dataset are formulated to cover different types of queries: questions that require specific facts, questions with descriptive answers (which require explanations or descriptions), and procedural questions, that is, questions that require a series of instructions or steps as a response. For every context, no more than five questions are included. For every question, one answer

is given. Data is collected in a variety of ways and verified through a manual annotation process to ensure the accuracy and relevance of responses. Tables 2 and 3 illustrate the need for manual correction e.g., named entities are written with errors (Applea instead of Eplovog), the word order is incorrect, and there are some grammatical errors.

Table 2. An example of context correction in the SQuAD-sr dataset

SQuAD	SQuAD-sr	Corrected SQuAD-sr
Apple's Safari had its first beta release in January 2003; as of April 2011, it had a dominant share of Apple-based web browsing, accounting for just over 7% of the entire browser market.	Safari od Applea imao je svoju prvu beta verziju u januaru 2003. godine; od aprila 2011. godine, imao je dominantan udeo Apel-baziranog veb pregledanja, što je činilo nešto više od 7% cele tržište pregledača.	Prva beta verzija Eplovog Safari pregledača objavljena je u januaru 2003. godine; od aprila 2011. godine, imao je dominantan udeo Epl-baziranog veb pregledanja, što je činilo nešto više od 7% celog tržišta pregledača.

Table 3. An example of a question and answer correction in the SQuAD-sr dataset

SQuAD	SQuAD-sr	Corrected SQuAD-sr
When was the first beta release for Safari?	Kada je prvi beta izdanje za Safari?	Kada je objavljena prva beta verzija Safarija?
January 2003	Januaru 2003.	U januaru 2003.

Current activities are focused on utilizing *Gemini 2.0* and *ChatGPT4* for automated grammar correction before manual review, to reduce the workload required for human correction. This approach enhances efficiency by minimizing the need for extensive human intervention while ensuring high linguistic accuracy and coherence in the final output. The lack of manually annotated datasets in Serbian makes these types of datasets particularly important.

3. QA Dataset from an Encyclopedia Using LLMs

This section outlines a method for generating a QA dataset from a structured encyclopedia containing tagged headwords and key entities. The goal is to leverage LLMs to generate high-quality question-answer pairs, ensuring that the dataset is both linguistically diverse and contextually accurate. The approach involves preprocessing, contextual understanding, question generation, answer extraction, and quality assessment. It is a work in progress, where the data has mostly been prepared and preprocessed, while the question and answer generation is still in the testing phase.

3.1. Data Preparation and Preprocessing

The encyclopedia used in this study is "Sveznanje". A part of it, about 20%, is available online on Wiki Izvornik[22], but the rest needs to be prepared from the paper copy. To that end, scanning was organized in the University Library Svetozar Marković, but after OCR, manual correction was also required. The next step was manual annotation of headwords and key terms marked with boldface in the original text. The part from Wiki Izvornik has 10,162 entries, while the digitized part currently comprises about 6000 entries, with the remaining entries to be finished by the end of this year.

The encyclopedia entries were segmented into smaller units based on paragraphs and headwords to ensure contextual coherence. They were structured using XML tags for entries, headwords, and key entities, and were used as input for data extraction and preprocessing to standardize the format for downstream tasks. In addition to the pre-tagged entities, a NER model is applied to identify additional significant terms. Entity linking is performed using external knowledge bases such as Wikidata to establish entity relationships. Any redundant or irrelevant text is removed, and duplicate entity occurrences are merged to prevent inconsistencies in question generation. TESLA-Sveznanje-QA is being generated using prompt engineering, where the LLM is given an encyclopedic article as context and asked to generate a set of questions and answers.

3.2. Question Generation Using LLMs

To produce diverse and contextually relevant questions, we utilize an LLM-based approach, which involves fine-tuning or prompting an existing transformer model such as GPT-4, Gemini 2.0, but we also plan to test other models in the future, like T5 and FLAN-T5.

Carefully designed prompts are used to instruct the LLM to generate questions based on encyclopedia content. Various strategies are explored, including:

- extractive QA generation, where the model is prompted to generate fact-based questions whose answers are directly found in the text;
- paraphrased QA generation, where the model is encouraged to generate multiple phrasings of the same question to enhance dataset diversity;
- reasoning-based QA generation, where questions requiring inferencing or logical reasoning beyond direct sentence extraction are included to improve depth.

[22] https://sr.wikisource.org/wiki/Свезнање

For domain-specific texts, we plan to train and use a domain-specific LLM, where fine-tuning is performed using an existing corpus of high-quality QA pairs before generating new samples from the encyclopaedia.

3.3. Answer Extraction and Verification

Once questions are generated, answer extraction is performed using both automated and human-in-the-loop methods:

- extractive answer selection, designed for fact-based questions, with the LLM task to extract precise spans of text from the encyclopaedia;
- generative answering, for cases where extractive answers are insufficient, and the LLM should generate summaries or inferred responses.

The answer consistency check follows to ensure the accuracy. A secondary LLM pass is applied to verify that the generated answer is supported by the source text. Additionally, human annotators validate samples of generated QA pairs.

The finalized QA dataset is structured in a standardized format, typically JSON or TSV, with fields including:

```
{
  "context": "Text excerpt from encyclopedia",
  "question": "Generated question",
  "answer": "Extracted or generated answer",
  "answer_type": "Extractive / Generative",
  "headword": "Tagged headword",
  "entities": ["List of linked entities"]
}
```

For example:

```
{
  "context": "PROTON, jezgro vodonikovog → atoma, nosi
  izvestan naboj pozitivnog el. i ima 1847 puta veću
  masu od elektrona; u poslednje vreme se smatra da je
  složen (→ neutron); sastavni je deo atomskog jezgra
  svake vrste atoma.",
  "question": "Koliko puta je masa protona veća od mase
  elektrona?",
  "answer": "1847 puta",
  "answer_type": "Extractive",
  "headword": "PROTON",
  "entities": ["Q2294, Q9121, Q2225, Q2348"]
}
```

Quality assessment involves both automated metrics and human evaluation. The fluency of generated questions is assessed using language modelling metrics (perplexity and coherence). Cross-referencing with external knowledge bases should ensure that generated answers remain factually accurate. A set of annotators also assesses the dataset for readability, correctness, and ambiguity.

The presented method is an efficient approach to generating a QA dataset from an encyclopaedia using LLMs. The proposed pipeline ensures contextual accuracy, question diversity, and answer reliability. Future research will focus on incorporating multi-turn QA generation, adversarial question generation for robustness testing, and multimodal QA datasets integrating images and structured data.

4. Textbooks

The third dataset that is being developed, *TESLA-domain-QA*, will primarily focus on environmental protection, computer science, mathematics, mining, geology, and energy. It will consist of at least 5,000 questions along with corresponding answers and contextual passages extracted from textbooks published by the Faculty of Mining and Geology.

This dataset is designed to ensure high domain specificity and reliability, as the source materials (textbooks) provide well-structured and authoritative information. The selection of these fields reflects their growing importance in contemporary research and industry, making the dataset valuable for training and evaluating domain-specific question-answering models.

Table 4 presents an example of *context* in Serbian with three question-answer pairs. Translation to English is not a part of the dataset, and it is presented here for readability.

The inclusion of well-defined contexts ensures that the dataset aligns with real-world educational materials, enabling the development of AI models capable of answering questions based on structured academic knowledge. Additionally, this dataset can contribute to educational applications, intelligent tutoring systems, and specialized question-answering models in Serbian.

Table 4. An example generated from textbook

Context	Question	Answer
Koreni rudarstva na tlu današnje Srbije nalaze se u praistorijskom periodu. Arheološka istraživanja potvrđuju da je jedan od naših najstarijih rudnika metala, za koji se danas zna, rudnik bakra Rudna glava kod Majdanpeka. Otkriveno je i istraženo više okana i kanala odakle potiče obiman fond nalaza: keramički sudovi, kameni batovi–obluci različitih veličina i namena, koštani alati i dr.	U kom periodu se nalaze koreni rudarstva, na tlu današnje Srbije?	U praistorijskom periodu.
	Koji je jedan od naših najstarijih rudnika metala?	Rudnik bakra Rudna glava kod Majdanpeka.
	Koji su neki od pronalazaka na području rudnika Rudna glava?	Keramički sudovi, kameni batovi obluci različitih veličina i namena, koštani alati i dr.
The roots of mining on the territory of present-day Serbia date back to prehistoric times. Archaeological research confirms that one of our oldest known metal mines is the copper mine at Rudna Glava, near Majdanpek. Several shafts and tunnels have been discovered and studied, yielding a significant collection of artifacts, including ceramic vessels, stone hammers of various sizes and functions, bone tools, and other items.	In which period do the roots of mining on the territory of present-day Serbia date back to?	Prehistoric times.
	What is one of our oldest metal mines?	The copper mine at Rudna Glava, near Majdanpek.
	What are some of the discoveries in the area of the Rudna Glava mine?	Ceramic vessels, stone hammers of various sizes and functions, bone tools, and other items.

5. Wikidata

The fourth dataset, *TESLA-Wikidata-QA,* is also currently under construction, but the core components are already designed. The core of the *TESLA-Wikidata-QA* is GeminiKnowledge-sr[23], a large-scale, structured QA dataset in Serbian, comprising over 127,000 question-answer pairs. The dataset spans over a broad spectrum of general knowledge and is uniquely organized through a taxonomy, a three-layer hierarchical structure of 'Area' (26), 'Topic' (317), and 'Facet' (6376) (e.g., 'Ekologija i održivost' ('Ecology and sustainability') → 'Biodiverzitet' (Biodiversity) → 'Tipovi ekosistema' ('Types of ecosystems')). Linking the hierarchical structure with Wikidata is followed by the dataset generation as explained in this

[23] https://huggingface.co/datasets/jerteh/GeminiKnowledge-sr

section and in (Stanković et al. 2025). This granular organization facilitates fine-grained exploration, enables targeted training of models on specific knowledge domains, and allows for nuanced evaluation of model performance across different levels of specificity. Created to address the significant lack of publicly available, high-quality Serbian language resources, GeminiKnowledge-sr is provided in JSONL format to encourage its widespread adoption and utilization within the research community.

TESLA-Wikidata-QA will consist of automatically generated contexts derived from the content of the Wikidata knowledge base, aligned with GeminiKnowledge-sr. Unlike the previous datasets, which rely on manually curated or translated texts, this dataset relies on structured knowledge from Wikidata to create a question-answering (QA) dataset in a fully automated manner. Such an approach produces a scalable, diverse, and up-to-date resource for training and evaluating QA models.

The process of generating a QA dataset from Wikidata involves multiple steps, including data extraction, context generation, question formulation, answer extraction, and quality assurance. Wikidata is a structured knowledge base that stores factual information in the form of triples:

```
(subject) - [predicate] → (object)
```

For example:

```
(Nikola Tesla) - [occupation] → (inventor)
(Nikola Tesla) - [occupation] → (physicist)
(Nikola Tesla) - [field of work] → (electrical engineering)
(Nikola Tesla) - [notable invention]→ (alternating current)
(Nikola Tesla) - [place of birth]→ (Smiljan, Austrian Empire)
(Nikola Tesla) - [known for]→ (Induction motor)
```

To extract relevant data, we use SPARQL queries to retrieve entities, properties, and relations related to target domains. The selection of entities and relations is guided by predefined topics, mentioned in the previous section. An example of SPARQL query to retrieve information about scientists and their fields of expertise is:

```
SELECT ?scientist ?scientistLabel ?field ?fieldLabel
WHERE {
  ?scientist wdt:P106 wd:Q901.  # Occupation: Scientist
  ?scientist wdt:P101 ?field.   # Field of work
  SERVICE wikibase:label { bd:serviceParam wikibase:language "sr". }
}
```

This query retrieves scientists and their respective fields of work, which can be used to generate contextual passages. We use Natural Language Generation (NLG) techniques to transform structured triples into coherent text passages. This transformation can be performed using predefined templates, such as:

- "[Subject] is known for [predicate] [object]."
- "[Subject] has made contributions to [predicate] in the field of [object]."

Alternatively, LLMs (e.g., GPT, T5) can be fine-tuned to generate more natural and varied contexts from structured data, but we are also experimenting with Gemini 2.0 for this task to obtain a greater variety of sentences and a more natural dataset.

Another example of using Natural Language Generation (NLG), where we transform Wikidata triples into a readable paragraph, is:

Extracted Data: *(Nikola Tesla) – [Known for] → (Alternating current, Induction motor, Wireless energy transmission).*

Generated Context: *"Nikola Tesla was a Serbian-American inventor, electrical engineer, and futurist known for his pioneering work in alternating current (AC) power transmission, the development of the induction motor, and early experiments in wireless energy transmission. His innovations in electrical engineering laid the foundation for modern power grids and influenced numerous technological advancements."*

Once the contexts are generated, questions are formulated based on the extracted information. The first approach for this step is template-based question generation, where predefined question templates are used to create structured questions:

- Who discovered [object]? → (Who discovered radium?)
- What is [subject] known for? → (What is Marie Curie known for?)

Another approach is LLM-Based Question Generation, where a fine-tuned question-generation model (such as T5) is employed to generate diverse and natural-sounding questions. The model is trained on existing QA datasets to learn how to rephrase and generate contextually appropriate questions.

Example of input-output:

Input (context from Wikidata):

"Marie Curie discovered radium and polonium, contributing significantly to the field of radioactivity."

Output (generated questions):

- *What elements did Marie Curie discover?*
- *Who contributed to the field of radioactivity?*

Answers are extracted directly from structured data in Wikidata. For example, if the question is *"Who discovered radium?"*, the corresponding Wikidata entity *(Marie Curie)* is retrieved as the answer.

There are two main strategies: 1) the answer is directly retrieved from the context, ensuring alignment with the original data; 2) if a simple answer does not fully capture the information, a generative approach (e.g., LLMs) is used to produce a more descriptive response.

We will illustrate here different types of generated questions:

- Fact-based: What is Nikola Tesla known for?
- Paraphrased: Which key inventions are attributed to Nikola Tesla?
- Specific: Who developed the alternating current system?
- Causal: Why is Nikola Tesla considered the pioneer in electrical engineering?

Using LLMs, the system can generate multiple phrasings of the same question, increasing dataset diversity.

An example of extractive vs. generative answers is:

- Extractive: "Marie Curie."
- Generative: "Marie Curie, a physicist and chemist, discovered radium along with polonium in her groundbreaking research on radioactivity."

More examples:

Question	Extracted Answer	Answer Type
Po čemu je poznat Nikola Tesla?	Naizmenična struja, indukcioni motor, bežični prenos energije.	Extractive
What is Nikola Tesla known for?	Alternating current, induction motor, wireless energy transmission.	
Gde je rođen Nikola Tesla?	Smiljan.	Extractive
Where was Nikola Tesla born?	Smiljan.	
Kakav je uticaj Nikola Tesla imao na elektrotehniku?	Razvio je sistem naizmenične struje, koji je postao osnova savremene distribucije električne energije.	Generative
What impact did Nikola Tesla have on electrical engineering?	He developed the alternating current system, which became the foundation of modern electricity distribution.	

To ensure accuracy and coherence, multiple automated and manual validation techniques are applied:

- fact verification, where answers are cross-checked with Wikidata and other knowledge sources,
- linguistic quality check, where generated contexts and questions are reviewed using grammar correction models (e.g., Gemini 2.0) before manual revision,
- diversity enhancement, where redundant or highly similar questions/ answers are removed to ensure dataset variety,
- human annotation (sample review), where a subset of generated QA pairs is manually reviewed for clarity, factual correctness, and naturalness.

The final dataset is stored in a structured format, such as JSON or TSV, with the following fields:

```
{
  "context": "Marie Curie discovered radium and
  polonium, contributing to radioactivity research.",
  "question": "What elements did Marie Curie discover?",
  "answer": "Radium and Polonium",
  "answer_type": "Extractive",
  "source": "Wikidata"
}
```

This format ensures compatibility with machine learning models and ease of use for training and evaluation tasks. The automatic generation of a QA dataset from Wikidata provides a scalable and up-to-date approach to dataset creation, leveraging structured knowledge to produce high-quality questions and answers. Future improvements will include enhancing paraphrasing techniques to create more natural questions and expanding the dataset to include reasoning questions using entity graphs. Previous examples can be represented as follows:

```
{
  "context": "Nikola Tesla was a Serbian-American
  inventor, electrical engineer, and futurist known for
  his pioneering work in alternating current (AC) power
  transmission, the development of the induction motor,
  and early experiments in wireless energy transmission."
  ,
  "question": "What is Nikola Tesla known for?",
  "answer": "Alternating current, induction motor,
  wireless energy transmission.",
  "answer_type": "Extractive",
  "source": "Wikidata"
}
```

6. Conclusion and Future Work

The application of this dataset, comprising four subsets, can be of substantial importance in various fields, including educational technologies, digital assistants, and information retrieval systems. By providing structured question-answer pairs, this dataset can serve as a valuable resource for improving machine comprehension and automated question-answering systems in Serbian.

Despite advancements in LLMs, even the most sophisticated models still lag behind human-level performance, highlighting the need for further development in the field of Natural Language Processing (NLP). The ability of AI systems to understand, reason, and generate linguistically and contextually appropriate responses remains an open challenge, particularly for languages with limited annotated resources.

A major issue in Serbian NLP research is the lack of manually annotated datasets, which restricts the ability to fine-tune models effectively. In this context, the dataset developed in this study can play a crucial role by bridging this gap and enabling the training of more accurate and context-aware AI models. By providing a high-quality, domain-relevant QA dataset, this research supports the development of Serbian language technologies, fostering improvements in machine translation, conversational AI, and educational tools.

Furthermore, this dataset may serve as a foundation for future advancements in language modeling, knowledge graph integration, and reasoning-based QA systems. We anticipate that this contribution will encourage further research and innovation in the domain of Serbian NLP, leading to the creation of more sophisticated AI-driven applications that can enhance human-computer interaction and knowledge accessibility.

Future work will focus on expanding the developed dataset by incorporating multimodal content (text, images, structured data), improving contextual understanding, and developing multilingual question-answering frameworks that enhance AI's ability to operate in diverse linguistic environments. By continuously refining and expanding this dataset, we aim to contribute to the evolution of Serbian AI systems, making them more robust, intelligent, and applicable to real-world scenarios.

Acknowledgment: This research was supported by the Science Fund of the Republic of Serbia, #7276, Text Embeddings - Serbian Language Applications – TESLA.

References

[34] Rogers, A., Gardner, M., & Augenstein, I. (2023). QA dataset explosion: A taxonomy of NLP resources for question answering and reading comprehension. *ACM Computing Surveys, 55*(10), 1–45. https://arxiv.org/pdf/2107.12708

[35] Rajpurkar, P., Jia, R., & Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. *arXiv preprint arXiv:1806.03822*. https://rajpurkar.github.io/SQuAD-explorer/

[36] Cenić, A. B., & Stojković, S. (2023). A Serbian question answering dataset created by using the web scraping technique. In *2023 58th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST)* (pp. 147–150). IEEE. https://doi.org/10.1109/ICEST58410.2023.10187370

[37] Chen, A., Stanovsky, G., Singh, S., & Gardner, M. (2019, November). Evaluating question answering evaluation. In Proceedings of the 2nd workshop on machine reading for question answering (pp. 119-124).

[38] Cvetanović, A., and Tadić, P. 2023. "Synthetic Dataset Creation and Fine-Tuning of Transformer Models for Question Answering in Serbian." *In 2023 31st Telecommunications Forum (TELFOR)*, 1–4. IEEE.

[39] Cvetanović, A., & Tadić, P. (2024). Synthetic dataset creation and fine-tuning of transformer models for question answering in Serbian. *arXiv preprint arXiv:2404.08617*. https://arxiv.org/html/2404.08617v1, https://paperswithcode.com/paper/synthetic-dataset-creation-and-fine-tuning-of

[40] Han, K., Ferreira, T. C., & Gardent, C. (2022). Generating questions from Wikidata triples. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 277–290). European Language Resources Association. https://aclanthology.org/2022.lrec-1.29/

[41] Hopkins, M., Le Bras, R., Petrescu-Prahova, C., Stanovsky, G., Hajishirzi, H., & Koncel-Kedziorski, R. (2019, June). SemEval-2019 task 10: Math question answering. In Proceedings of the 13th International Workshop on Semantic Evaluation (pp. 893-899).

[42] Korablinov, V., & Braslavski, P. (2020). RuBQ: A Russian dataset for question answering over Wikidata. In: Pan, J.Z., et al (Eds.), *The Semantic Web – ISWC 2020: 19th International Semantic Web Conference, Athens, Greece, November 2–6, 2020, Proceedings, Part II* (Vol. 12507, pp. 97–110). Springer. https://doi.org/10.1007/978-3-030-62466-8_7

[43] Stanković, R., Janković, N., Rađenović, J., Ikonić-Nešić, M. (2025) From LLM Generation to Knowledge Representation: Creating and Structuring the GeminiKnowledge--sr QA Dataset for Serbian, *GOBLIN 2025: 1st GOBLIN*

Workshop on Knowledge Graph Technologies, Leipzig, Germany, https://www.dbpedia.org/events/goblin25-workshop/ (submitted).

[44] Usbeck, R., Yan, X., Perevalov, A., Jiang, L., Schulz, J., Kraft, A., Möller, C., et al. (2024). QALD-10–The 10th challenge on question answering over linked data: Shifting from DBpedia to Wikidata as a KG for KGQA. *Semantic Web,* 15(6), 2193–2207. https://doi.org/10.3233/SW-233471

[45] Zaib, M., Zhang, W. E., Sheng, Q. Z., Mahmood, A., & Zhang, Y. (2022). Conversational question answering: A survey. *Knowledge and Information Systems, 64*(12), 3151–3195. https://doi.org/10.1007/s10115-022-01744-y

Kreiranje skupa za obučavanje modela za odgovaranje na pitanja na srpskom jeziku

Ranka Stanković, Jovana Rađenović, Maja Ristić, Dragan Stankov

Sažetak

Razvoj i primena veštačke inteligencije u jezičkim tehnologijama značajno su napredovali poslednjih godina, posebno u domenu zadatka odgovaranja na pitanja (Question Answering - QA). Dok su postojeći resursi za QA zadatke razvijeni za glavne svetske jezike, srpski jezik je relativno zanemaren u ovoj oblasti. Ovaj rad predstavlja inicijativu za kreiranje obimnog i raznovrsnog skupa podataka za obučavanje modela za odgovaranje na pitanja na srpskom jeziku, koji će doprineti unapređenju jezičkih tehnologija za srpski jezik.

Pored brojnih istraživanja o jezičkim modelima u poslednjih nekoliko godina, mnogo je urađeno i na referentnim skupovima podataka potrebnim za praćenje napretka modeliranja. Posebno je puno urađeno kada je reč o odgovaranju na pitanja i razumevanju pročitanog mada, uglavnom kada je reč o velikim jezicima (Rogers et al. 2023). U radu se pruža pregled različitih formata i domena raspoloživih višejezičnih i jednojezičnih resursa, sa posebnim osvrtom na srpski jezik (Cenić & Stojković 2023; Cvetanović & Tadić 2024).

U okviru projekta TESLA (Text Embeddings - Serbian Language Applications) radi se na pripremi skupa podataka: kontekst, pitanja i odgovori, prikupljenih iz različitih domena. Skup će biti sačinjen od četiri manja podskupa. U cilju izrade prvog podskupa, TESLA-SQuAD-sr, podskup Stanfordovog skupa SQuAD (Rajpurkar et al. 2018), gde je odgovor segment teksta, prevodi se i prilagođava, birajući teme kao što su:

Nikola Tesla, klimatske promene, građevina, geologija, itd. Podskup trenutno ima 5600, ali će imati oko 7000 pitanja sa pratećim odgovorima. Drugi podskup, TESLA-Sveznanje-QA, se priprema na osnovu enciklopedije Sveznanje na čijoj retrodigitalizaciji se intenzivno radi. Treći podskup, TESLA-domain-QA, koji se priprema na osnovu udžbenika u izdanju Rudarsko-geološkog fakultet, biće vezan pre svega za zaštitu životne sredine, informatiku i energetiku i sadržaće najmanje 5000 pitanja sa odgovorima i datim kontekstom ekscerpiranim iz udžbenika. Četvrti podskup, TESLA-Wikidata-QA, uglavnom će sadržati automatski generisane kontekste na osnovu sadržaja baze znanja Wikidata i Gemini 4.0 jezičkog modela.

Pitanja su pažljivo formulisana kako bi pokrila različite tipove upita: pitanja koja zahtevaju konkretne činjenice, pitanja sa deskriptivnim odgovorom (koja traže objašnjenja ili opis), i proceduralna pitanja, odnosno pitanja koja kao odgovor zahtevaju niz uputstava ili koraka. Podaci se prikupljaju na različite načine i verifikuju kroz proces ručnog anotiranja kako bi se obezbedila tačnost i relevantnost odgovora. Nedostatak ručno anotiranih skupova podataka na srpskom jeziku čini doprinos ovog istraživanja posebno značajnim.

Zaključak rada ukazuje na značaj i potencijal primene ovog skupa podataka u različitim oblastima, uključujući obrazovne tehnologije, digitalne asistente, i sisteme za pretragu informacija. Predstavljeni rezultati doprinose unapređenju jezičkih tehnologija za srpski jezik, i nadamo se da će podstaći dalja istraživanja i razvoj u ovoj oblasti.

Ključne reči: veštačka inteligencija, obrada prirodnog jezika, jezički resursi, anotirani skupovi, ekstrakcija informacija, odgovaranje na pitanja