SrpCNNeL: Serbian Model for Named Entity Linking

Milica Ikonić Nešić, Saša Petalinkar, Ranka Stanković, Miloš Utvić, Olivera Kitanović

SrpCNNeL: Serbian Model for Named Entity Linking | Milica Ikonić Nešić, Saša Petalinkar, Ranka Stanković, Miloš Utvić, Olivera Kitanović | Annals of Computer Science and Information Systems | 2024 | |

10.15439/2024F8827

http://dr.rgf.bg.ac.rs/s/repo/item/0009165

Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 465–473
DOI: 10.15439/2024F8827
ISSN 2300-5963 ACSIS, Vol. 39

SrpCNNeL: Serbian Model for Named Entity Linking

Milica Ikonić Nešić*, Saša Petalinkar†, Ranka Stanković‡, Miloš Utvić* and Olivera Kitanović‡
*University of Belgrade, Faculty of Philology, Belgrade, Serbia
Email: milica.ikonic.nesic@fil.bg.ac.rs, milos.utvic@fil.bg.ac.rs
†University of Belgrade, Belgrade, Serbia
Email: sasa5linkAr@gmail.com
‡University of Belgrade, Faculty of Mining and Geology, Belgrade, Serbia
Email: ranka.stankovic@rgf.bg.ac.rs, olivera.kitanovic@rgf.bg.ac.rs

***Abstract*—This paper presents the development of a Named Entity Linking (NEL) model to the Wikidata knowledge base for the Serbian language, named SrpCNNeL. The model was trained to recognize and link seven different named entity types (persons, locations, organizations, professions, events, demonyms, and works of art) on a dataset containing sentences from novels, legal documents, as well as sentences generated from the Wikidata knowledge base and the Leximirka lexical database. The resulting model demonstrated robust performance, achieving an F1 score of 0.8 on the test set. Considering that the dataset contains the highest number of locations linked to the knowledge base, an evaluation was conducted on an independent dataset and compared to the baseline Spacy Entity Linker for locations only.**

I. Introduction

NAMED Entity Linking (NEL) is one of the important tasks in Natural Language Processing (NLP) that focuses on identifying and disambiguating entities mentioned in the text by linking them to a corresponding knowledge base [1]. This process is essential for structuring unstructured data, which is crucial for various NLP applications such as information retrieval, sentiment analysis, and knowledge base population. NEL is particularly significant for low-resource, morphologically rich languages like Serbian due to the unique challenges and benefits it presents [2], [3].

During the last decade, various works have addressed named entity linking, recognition, and disambiguation. The task of recognizing mentions of entities in text and disambiguating them to the corresponding entities in a knowledge base (KB) is called Entity Linking (EL)[4]. EL systems have demonstrated remarkable performance on standard benchmarks, a feature largely attributable to the advent of contemporary language models[5]. It has found innumerable applications in a wide range of downstream tasks, such as Question Answering [6], Information Extraction [7], Historic Newspaper optical character recognition (OCR) [8], Esports News [9], EL for Tweets [10], and Biomedical areas [11], [12]. The importance of NEL extends beyond entity recognition; it plays a crucial role in organizing and extracting meaningful information from large text corpora. Effective NEL can significantly improve the accuracy of downstream NLP tasks by providing a structured understanding of entities and their relationships within the text. This is especially important for processing large volumes of unstructured data, which is common in many real-world applications. For instance, NEL can improve information retrieval by linking query terms to relevant entities, enhance sentiment analysis by accurately identifying sentiment targets, and support KB population by automatically updating entity information.

New research focuses on a survey of the scientific literature on NEL, including named entity recognition and disambiguation, covering 200 works by focusing on 43 papers (5 surveys and 38 research works). The authors also described and classified 56 resources, including 25 tools and 31 corpora for English and other languages [13]. Balog [1] defined the problems of EL, NER, and NED (Named Entity Disambiguation) by presenting the general process of Entity Linking (EL), which consists solely of NER and NED. He asserts that named entities must be extracted before they can be disambiguated.

ScispaCy [14] integrates spaCy for biomedical text processing, offering efficient and accurate entity linking to knowledge bases like the Unified Medical Language System (UMLS)[15] and Gene Ontology[16]. These models demonstrate significant improvements in linking biomedical entities, highlighting spaCy's versatility in handling domain-specific NLP tasks. Radboud Entity Linker (REL) [17] is an open-source toolkit for entity linking, which builds on neural components from natural language processing research and is provided as a Python package and a web API.

The linking of named entities has sharply increased over the years for multilingual models [18], [19], [20]. Bi-encoder Entity Linking Architecture (BELA) presents the first fully end-to-end multilingual entity linking model that efficiently detects and links entities in texts in any of 97 languages [21]. Additionally, there has been a growing body of research focused on enhancing neural models by leveraging relational knowledge from semantic networks using Wikidata [22], [23], [24]. Linking entities to the Wikidata knowledge base is currently a highly relevant topic, and one of the projects from 2021 is the Spacy Entity Linker (spacy-entity-linker 1.0.3) pipeline for spaCy that performs linked entity extraction with Wikidata, which can be used as a multilingual entity model for linking with the Wikidata knowledge base. The University

Library of Mannheim (Universitätsbibliothek Mannheim, abbreviated UB Mannheim) developed spaCyOpenTapioca for the task of linking named entities to concepts (items) in Wikidata in spaCy using OpenTapioca [25]. This system achieved an F1 score of 0.09 on an Italian-Serbian corpus of 10,000 aligned segments (sentences) taken from different novels, named the It-Sr-NER corpus [26], [27]. Early research leveraged word2vec and convolutional neural networks (CNN) to capture the correlation between mentioned context and entity information [28], [29] and link entities for languages such as Chinese [30] and Italian [31]. An end-to-end entity linking system for the Greek language was developed in order to extend the Radboud Entity Linker (REL) toolkit to support modern Greek. The authors investigate three different mention detection approaches using spaCy, Flair, and BERT [32].

However, one common issue with current EL approaches is that they require massive amounts of training data. Consequently, training models for named entity recognition and linking with knowledge bases for low-resource languages is a highly challenging endeavor. In the case of the Serbian language, the problem encountered when applying multilingual models for linking recognized entities to the Wikidata knowledge base is the inability to recognize inflected forms of entities. Resolving this problem was one of the main motivations for this research.

Serbian, with its complex morphological structures and limited NLP resources, poses significant hurdles for accurate named entity recognition and linking. The intricate morphology, including rich inflectional patterns, necessitates advanced models capable of handling various word forms and syntactic nuances. Implementing NEL for Serbian using spaCy provides a robust framework to address these challenges by leveraging advanced machine-learning techniques and pre-trained language models tailored for low-resource settings. Considering the challenges presented by the Serbian language, this work represents, to the best of our knowledge, one of the first attempts to train a model for named entity recognition and linking to the corresponding items in Wikidata, with a primary focus on locations using the previously trained SrpCNNER2 model for the NER task. This model was trained on a dataset containing Serbian novels published between 1840 and 1920 [33], publicly available newspaper articles, and sentences generated for the NER task from the Wikidata [34] knowledge base and Leximirka lexical database [35], achieving an F1 score of approximately 0.71 on the test dataset.

In conclusion, implementing named entity linking for Serbian using spaCy addresses the specific linguistic challenges posed by the language's morphology and enhances the overall quality and usability of NLP applications in low-resource settings. This research underscores the importance of developing tailored NLP solutions that cater to the unique needs of morphologically rich languages, ultimately contributing to more inclusive and comprehensive language technologies. The paper is organized into several sections as follows. Section II briefly presents the process of preparing the dataset for the training model. Data conversion and the development of a knowledge base, which enabled the training of a CNN-based NEL model for Serbian, named SrpCNNeL, are presented in Section III. Evaluation of the model in two different settings: the first discusses the model's performance on the test subset of the prepared dataset, and the second carries out a detailed evaluation on novel and newspaper articles that were not present in the training dataset, can be found in Section IV. Finally, conclusions and plans for future work are presented in Section V.

II. Data Preparation

One of the key issues addressed in this paper was overcoming the problem of linking inflected forms of entities mentioned in the text with the Wikidata KB. Since items in Wikidata have labels in the nominative form, such as *Beograd* (Q3711) (Belgrade), it was necessary to create a synthetic dataset with inflected forms to ensure that all inflected forms of the mentioned entity in the text are linked with the same item in the KB. For instance, *Beograda*, *Beogradu*, or *Beogradom* are linked to the item *Beograd* (Q3711).

Therefore, for the purposes of this research, the training dataset was created and evaluated in two stages. In the first stage, the synthetic training dataset was generated using data from the Wikidata knowledge base and the Leximirka lexical database [35].

Given that the lexical database Leximirka has semantic markers such as *NPropN* for named entities, as well as **Dr** for country, **Gr** for city, **Hum** for person, **Oro** for mountain, **Hyd** for river, and **Org** for organizations, it was possible to access named entities that belong to any of the mentioned categories. On the other hand, using the Wikidata Query Service and constructing a SPARQL query [36], it was possible to create a list of all items belonging to the categories of country, city, mountain, river, or organization which have labels in the Serbian language. This was achieved by using specific properties in Wikidata for each category, as shown in Table I.

The SPARQL query for extracting instances of countries with appropriate QIDs from Wikidata can be retrieved using the following query:

```
# Countries with labels in Serbian
SELECT DISTINCT ?country ?cLabel
WHERE {
  ?country rdfs:label ?cLabel;
  #instance of country
  wdt:P31 wd:Q6256.
filter langMatches(lang(?cLabel),"sr")
}}
```

TABLE I
PROPERTIES FOR SPECIFIC CATEGORY OF NAMED ENTITES

semantic marker	category	instance of (P31)
Dr	country	wd:Q6256
Gr	sity	wd:Q515
Oro	mountain	wd:Q46831
Hyd	river	wd:Q4022

TABLE II
EXAMPLE OF SQL FUNCTIONS AND GENERATED SYNTHETIC SENTENCES

mark	num.	SQL function	synthetic sentences
Gr	4	select * from dbo.fnGenerisiReceniceNER(N'<s>Posetio sam <LOC>',N'</LOC> prošlog leta. </s>','**Gr**','4',100)	('Posetio sam **Beograd** prošlog leta. ' 'links': (12,19): 'Q3711':1.0,'entities':[(12,19,'LOC')])
Dr	2	select * from dbo.fnGenerisiReceniceNER(N'<s>U srcu <LOC>',N'</LOC> leži bogatstvo narodnih običaja. </s>','**Dr**','2',100)	('U srcu **Srbije** leži bogatstvo narodnih običaja. ' 'links': (7,13): 'Q403':1.0,'entities':[(7,13,'LOC')])
Hyd	2	select * from dbo.fnGenerisiReceniceNER(N'<s>Voda <LOC>',N'</LOC> je bistra. </s> ','**Hyd**','2',60)	('Voda **Dunava** je bistra. ' 'links': (5,11): 'Q1653':1.0,'entities':[(5,11,'LOC')])
Oro	7	select * from dbo.fnGenerisiReceniceNER(N'<s>Našli smo izvor vode na <LOC>',N'</LOC>. </s>','**Oro**','5',60)	('Našli smo izvor vode na **Durmitoru** . ' 'links': (24,33): 'Q212836':1.0,'entities':[(24,33,'LOC')])

After extracting QIDs from Wikidata, it was possible to input them into Leximirka and link them to the corresponding lexical units. Since Leximirka contains inflected forms of all lemmas, it was possible to derive the correct grammatical forms for all grammatical cases of named entities that appear in it and that are linked to Wikidata. A function was developed in MS SQL Server to generate synthetic sentences for a given sentence template. In Table II, examples of SQL queries and appropriately generated synthetic sentences are presented. The column "num" represents the ordinal number of the case in the Serbian language.

In the end, this dataset contains 16,869 sentences, including only those with locations (rivers, cities, countries, lakes, and mountains) and organizations.

After creating a synthetic dataset, the dataset was expanded by 22,730 sentences from various novels written or translated into Serbian, as well as legal documents. The novels whose sentences belong to this dataset include Jules Verne's *"Around the World in Eighty Days"*[37], Orwell's *"1984"*[38], and novels from the It-Sr-NER corpus [26]: Umberto Eco's *"The Name of the Rose"*, Carlo Collodi's *"The Adventures of Pinocchio"*, Elena Ferrante's *"Those Who Leave and Those Who Stay"*, and Luigi Pirandello's *"One, None and a Hundred Thousand"*. The corpus also includes five novels by Serbian writers: Ivo Andrić's *"Anikina vremena"* ("Legends of Anika") and *"Na drini ćuprija"* ("The Bridge on the Drina"), Borisav Stanković's *"Nečista krv"* ("Impure Blood"), as well as legal documents from *Intera* available on the Biblish digital library [39]. The training dataset was prepared through several steps. Initially, the dataset was annotated with named entities using the Jerteh-355-tesla [40] and the Jerteh-355 [41] fine-tuned language model for NER. Jerteh-355 is based on the RoBERTa-large architecture [42]. The model is trained to recognize seven categories: demonyms (DEMO), professions and titles (ROLE), works of art (WORK), person names (PERS), locations (LOC), events (EVENT), and organizations (ORG). After automatic annotation, the INCEpTION tool [43] was used for the manual correction and linking of named entities with the Wikidata knowledge base. An example of annotation and linking with Wikidata using the INCEpTION tool is presented in Fig. 1. The named entities were linked in an additional layer of annotation where a Wikidata identifier was assigned to each entity instance.

After the dataset was prepared, it contained 35,955 sentences in total. Among the annotated entities, as well as those linked to Wikidata, the majority were recognized as locations (LOC), of which only 322 were not linked to Wikidata (Fig. 2), primarily those locations for which corresponding items do not exist in Wikidata.

The distribution of all named entities by type, showing the proportion of those linked to Wikidata versus to those that are not, as well the number of all named entities of appropriate class, is illustrated in the Fig. 3 The Fig. 4 shows the percentage of unique entities relative to the total number of entities recorded in the dataset. Observing the further distribution of linked entities by the number of appearances in the dataset, Serbia (Q403) is the most frequently linked entity to Wikidata. Table III presents the ten entities that are most frequently linked to Wikidata. Although such a distribution is not optimal for training entity linking in a morphologically rich language, it is not uncommon for a corpus dominated by natural text. Fig. 5 illustrates the relationship between the number of entity occurrences and the frequency of entity repetitions. It can be observed that nearly 3,000 entities appear only once in the dataset, while only one entity (Serbia (Q403)) appears more than 700 times.

III. SRPCNNEL MODEL FOR SERBIAN

This section outlines the methodology employed to train the Entity Linker for the Serbian language using the spaCy framework. The process encompassed data preparation, the development of a knowledge base using Wikidata, the training of the entity linking model, and the evaluation of its performance.

TABLE III
TEN MOST COMMON LINKED ENTITIES IN DATASET

QID	Name	Count
Q403	Srbija	785
Q236	Crna Gora	403
Q11428966	Vinston Smit (1984)	327
Q37226	učitelj	277
Q838261	Savezna Republika Jugoslavija	248
Q2587533	Fileas Fog	243
Q127885	Srbi	239
Q327055	radnik	222
Q170287	Subotica	188
Q37024	Srbija i Crna Gora	183

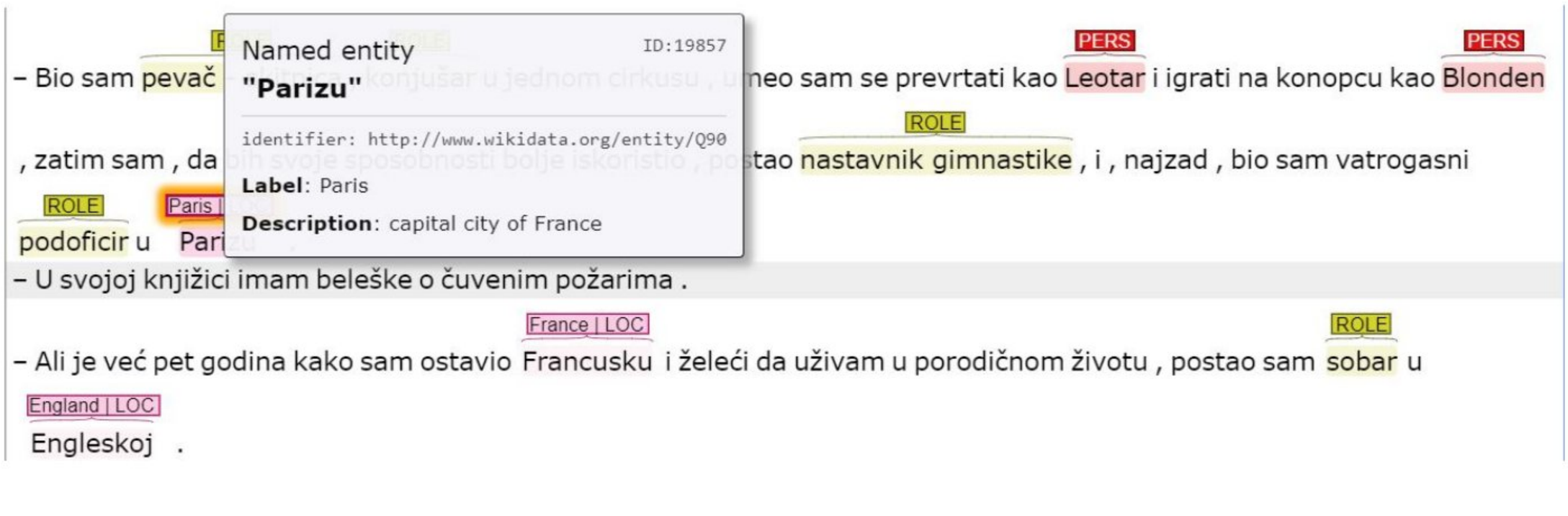

Fig. 1. An example of annotation in INCEpTION

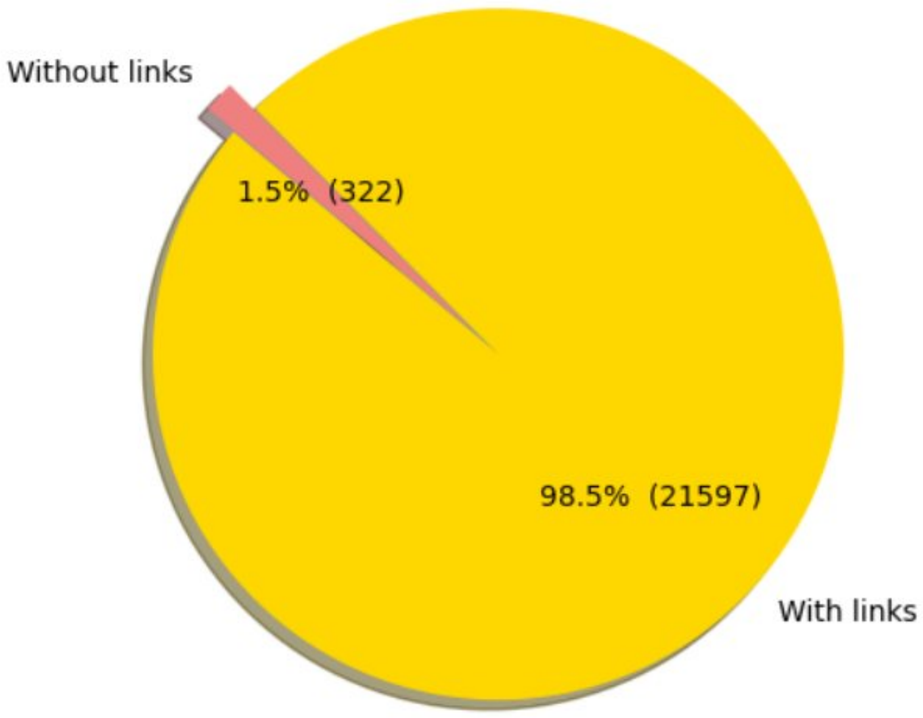

Fig. 2. Percentage of entities that are linked to the KB and those that are not

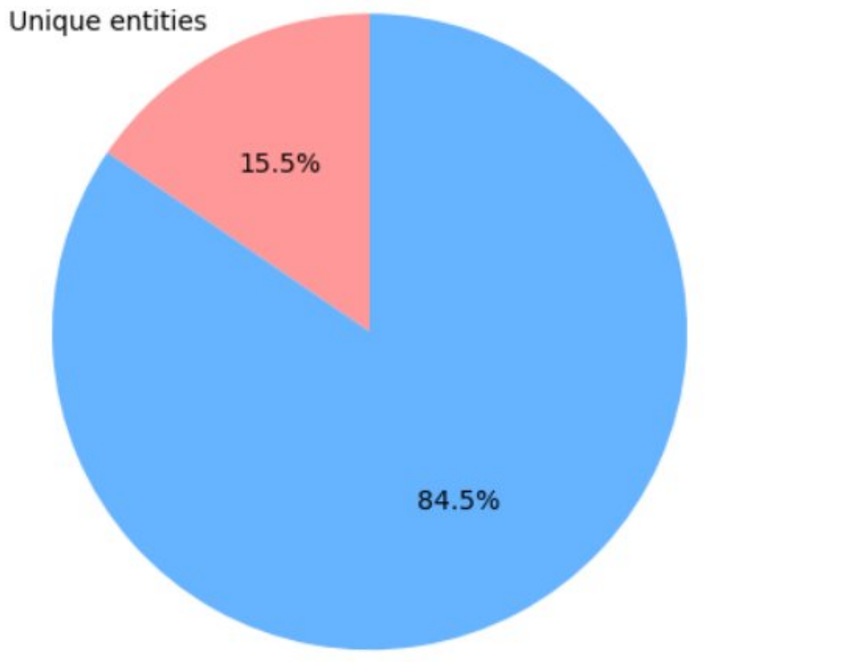

Fig. 4. Percent of unique linked entities in corpus

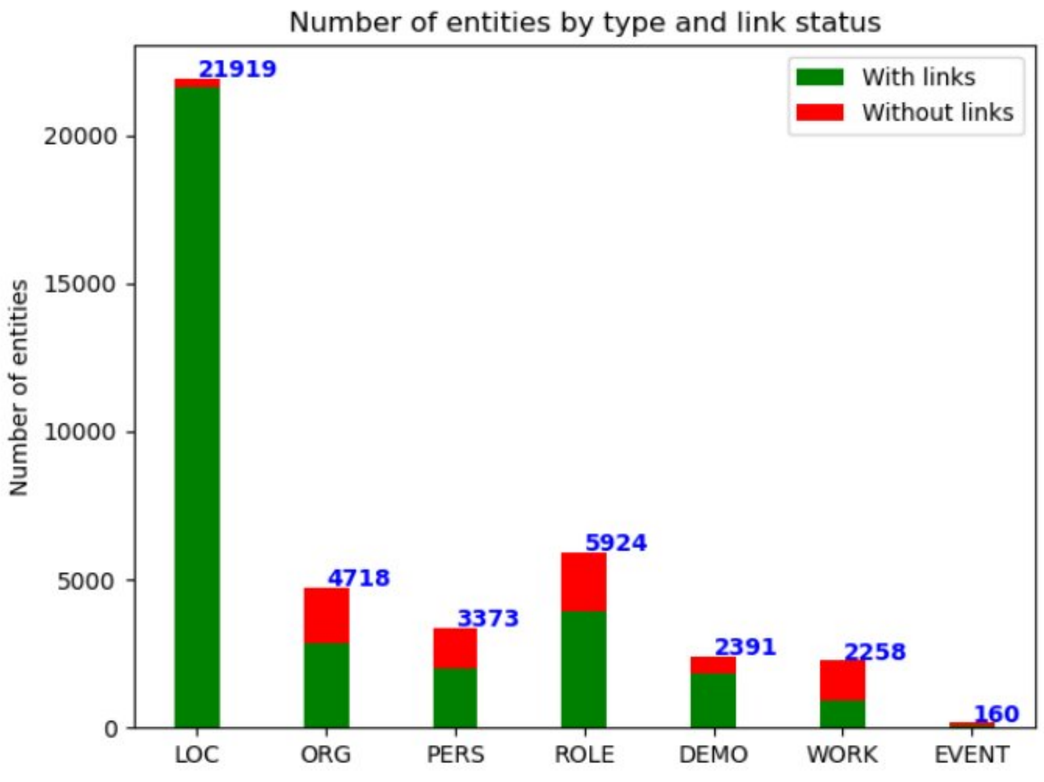

Fig. 3. Distribution of named entity types with and without linking to Wikidata

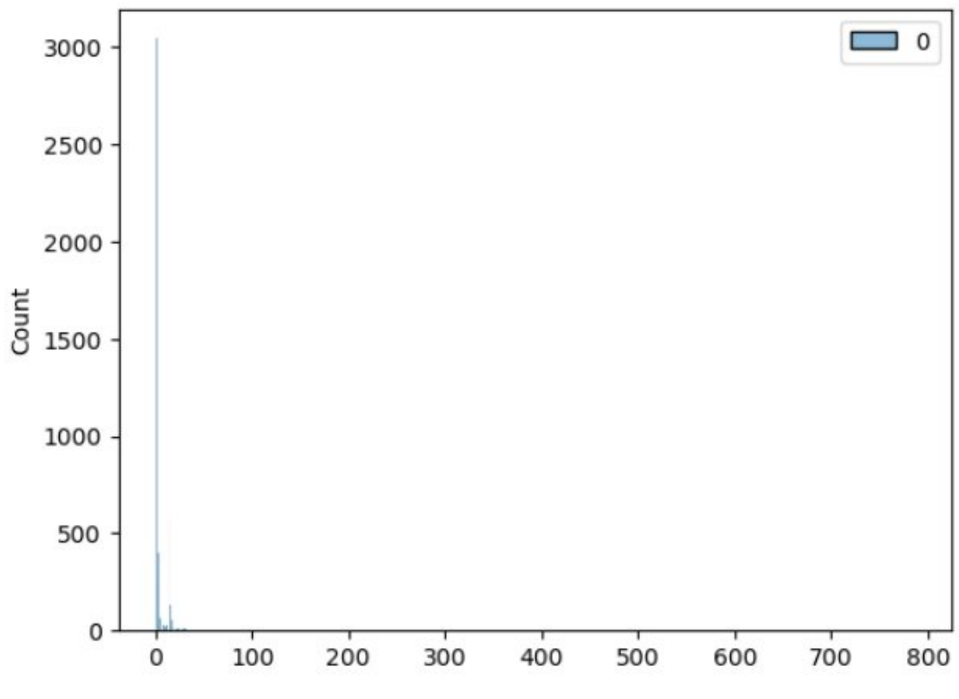

Fig. 5. Distribution of unique entities by the number of appearance in the dataset

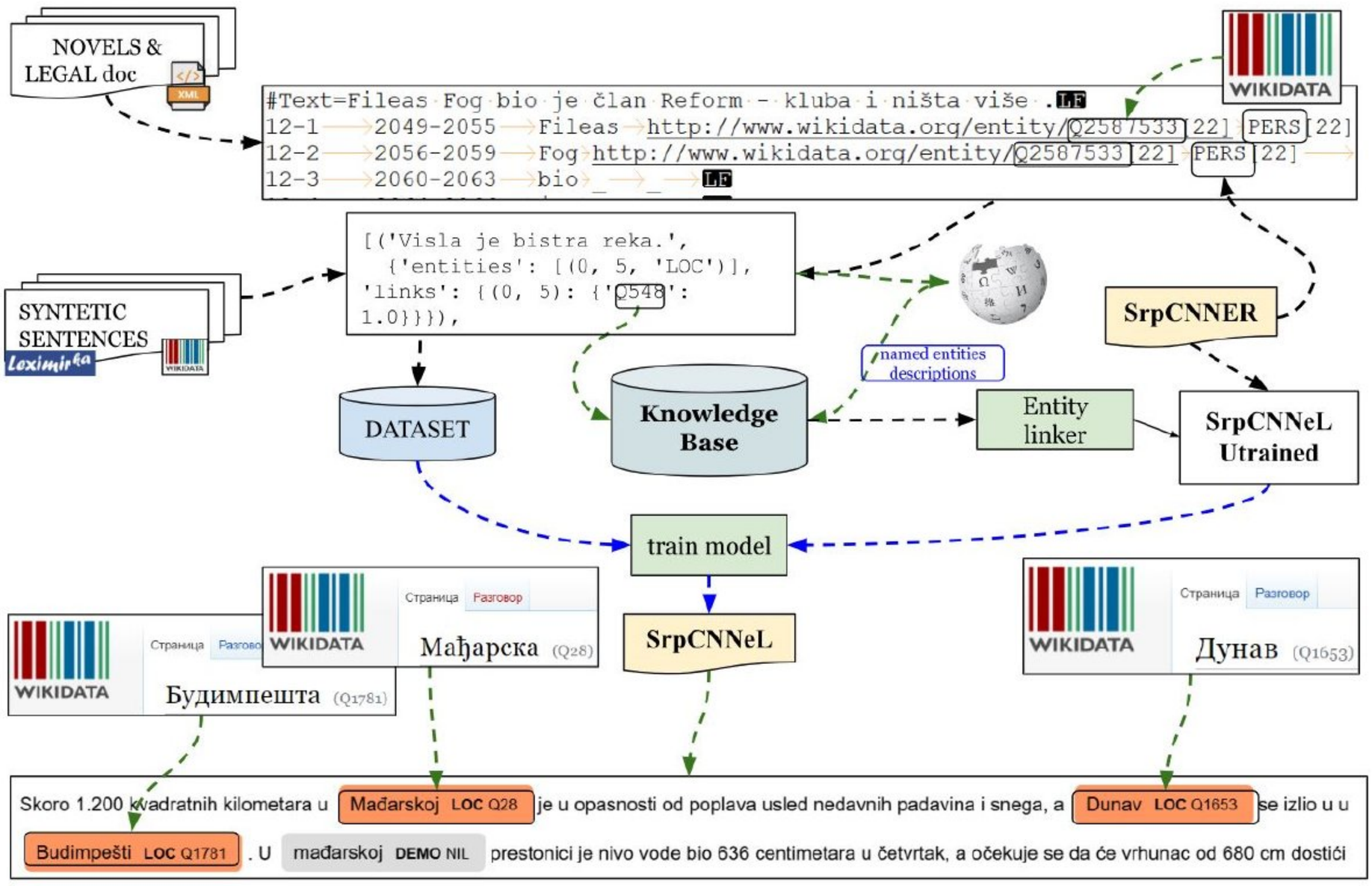

Fig. 6. The workflow of the whole process of NEL

The workflow of the previous process is presented in Fig. 6. evaluation of its performance.

A. Data Conversion

For training the NEL model, the training dataset contained 35,955 sentences in total. The research involved converting data from different formats into one usable by spaCy, with serialization into the `.pkl` (pickle) format. The pickle format is efficient for storage and quick loading during model training and inference. The code used for this research included functions to convert TSV and custom text formats into a spaCy-compatible format, ensuring accurate entity linking and recognition.

B. Knowledge Base Development

The knowledge base was developed from the extracted entities and appropriate QIDs in the dataset and enriched with descriptions obtained from Serbian Wikipedia pages. This involved extracting entities and their corresponding descriptions to form a comprehensive knowledge base. This step was particularly important for morphologically rich languages like Serbian, as it allowed for the capturing of all inflected forms of words that appear in the dataset.

Instead of relying on lemmatization, inflected forms of words were incorporated into the knowledge base. This decision stems from the complexity of lemmatizing multi-word named entities in Serbian, where a token-by-token lemmatization approach is inadequate and often incorrect.

Unique identifiers from Wikidata were extracted for each entity, along with all texts that correspond to each entity's aliases. Probabilities were not extracted from the dataset; instead, each alias was assigned a uniform starting probability. Descriptions from Wikipedia were connected to each unique identifier from Wikidata and vectorized using the base spaCy pipeline. These descriptions of entities, as well as all possible aliases, were then inserted into the knowledge base. Ensuring that inflected forms are included in the knowledge base facilitates their recognition, making it possible to accurately link entities in the text to their corresponding entries in the Wikidata knowledge base. This comprehensive approach ensures that the entities recognized in the text are accurately linked, enhancing the model's performance in handling the complex linguistic variations inherent in the Serbian language.

C. Training the Entity Linking Model

The SrpCNNeL model extends the pre-trained NER model for Serbian, SrpCNNER2, by integrating a spaCy entity linking layer, where the base model employed for entity linking is **spacy.EntityLinker.v2**. The SrpCNNER2 model is trained using the spaCy Python module, version 3.2, employing the same model architecture as SrpCNNER [44], using 148,819 sentences dataset, sourced from three distinct components: srpELTeC-gold-extended, newspaper articles and generated sentences. Importantly, the dataset includes sentences that do not contain any named entities. The breakdown is as follows:

- **srpELTeC-gold-extended**: This component contributes 54,423 sentences, encompassing 986,567 tokens and

35,772 named entities. This corpus is an extension of the srpELTeC-gold corpus [45], which contains sentences from old Serbian novels within the SrpELTeC corpus.

- **Newspaper articles**: This segment consists of 9,498 sentences, amounting to 235,953 tokens and 28,496 named entities.
- **Generated sentences**: This category comprises the largest portion, with 84,898 sentences, totaling 670,722 tokens and 85,642 named entities. Sentences are generated by the Wikidata knowledge base and sentences generated on the basis of Leximirka lexical database [46], in the form of CONLLu files which contain information about NER, POS-tag, and lemma.

SrpCNNER2 is trained to recognize seven categories of entities: persons (PERS), professions (ROLE), demonyms (DEMO), organizations (ORG), locations (LOC), artworks (WORK), and events (EVENT). The SrpCNNeL model was trained using the previously converted data in pkl format and the created knowledge base. Sentences were randomly shuffled and split into training and test sets with a ratio of 8:2, i.e., 31,679 sentences in training and 3,720 sentences in the test set. SpaCy's pipeline was employed to integrate the NER component with the *entity linker*. To feed training data into the entity linker, the pkl format presents a list of structured tuples. The first part is the raw text, and the second part is a dictionary of annotations. The dictionary defines the named entities we want to link ("entities"), as well as the actual gold-standard links ("links"). An example of such a tuple is the following:

('Visla je bistra reka.', 'links': (0, 5): 'Q548': 1.0, 'entities': [(0, 5, 'LOC')])

This integration was essential to ensure that entities identified in the text were correctly linked to their corresponding entries in the knowledge base. The dropout rate was set to 0.2, and the optimizer was inherited from SrpCNNER2. The training data, serialized into pickle files, was loaded, and the model was trained iteratively (with 100 iterations) using examples from the training dataset.

IV. Evaluation

Model evaluation performed on the previously discussed test set demonstrated the following performance metrics: a precision of 0.79, a recall of 0.83, and an F1 score of 0.80, indicating a robust model performance.

In addition to this, and in order to better understand how the model behaves at the sentence level, we introduce a new metric called *"accuracy by sentence"*. This measure provides a more holistic view of the model's performance by considering the correct prediction of entire sentences rather than individual entities. A sentence is considered correct if all predicted entities in it correspond to those annotated in the dataset, pointing to the same text and being linked identically. The type of entity was not considered, as the spaCy entity linker does not account for entity types. Accuracy by sentence is calculated according to Equation 1.

$$accuracy_{sentence} = \frac{correct_{sentence}}{total_{sentences}} \tag{1}$$

Introducing this metric is motivated by the need to evaluate the model's performance in practical, real-world scenarios, where accurate entity linking across entire sentences is crucial for applications such as information extraction, knowledge base population, and automated content analysis. By focusing on sentence-level accuracy, we can better assess the model's ability to understand and process context, ensuring that linked entities are not only identified correctly but also coherent within their respective sentences. The result on the test set is an accuracy by sentence of 0.67, as shown in Fig. 7. This highlights the model's capability to correctly link entities in a significant portion of the sentences, though there remains room for improvement to achieve higher accuracy in more complex or nuanced cases.

In this research, the main focus was on the recognition of linked locations and organizations, as the occurrence of names and roles depends on the text itself. For example, literary texts more frequently feature *characters* from novels or *historical persons*, while newspapers predominantly mention *politicians, athletes, actors*, etc. Even the roles differ, with newspapers mentioning *politicians, collaborators, directors*, while in the novels within this corpus, roles such as *duke, king, servant*, etc., are more prevalent. However, to highlight both the advantages and disadvantages of such a system, in this study, all named entities present in the knowledge base were linked within the corpus.

The performance achieved by the model indicates that a larger dataset is necessary for training the entity linking model in this case. The scatter plot presented in Fig. 8 suggests that the model has generally good classification performance, as

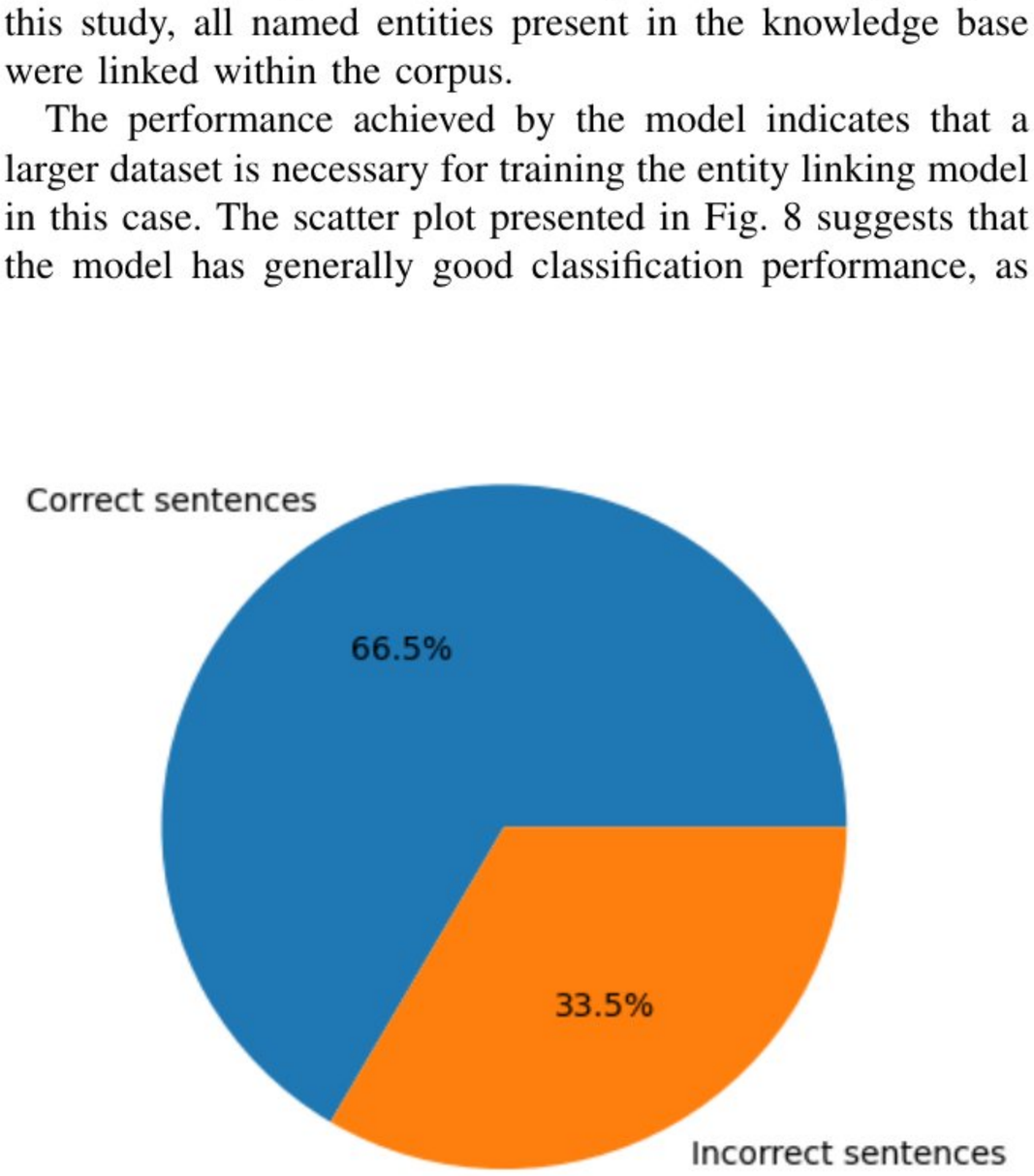

Fig. 7. Accuracy by sentence

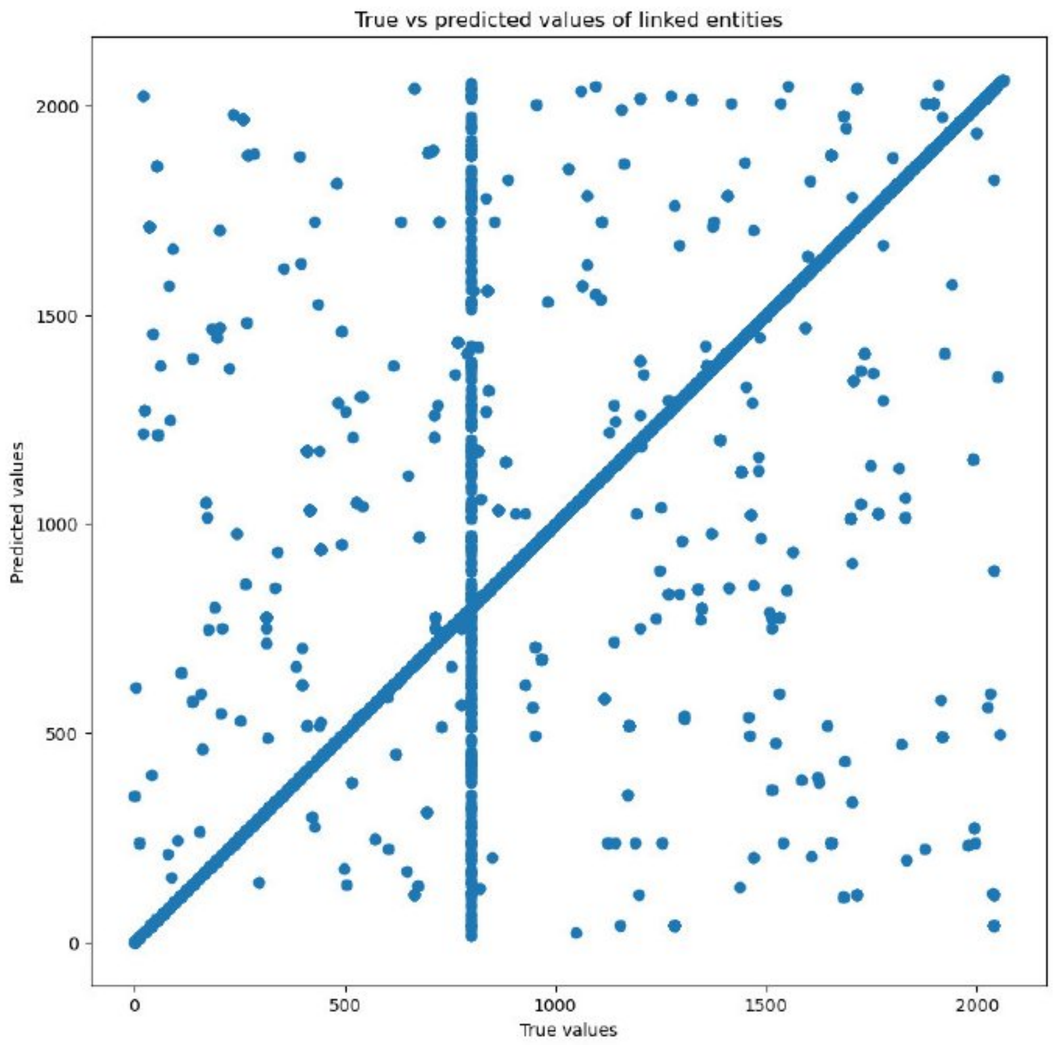

Fig. 8. Scatter plot by Wikidata links

indicated by the concentration of points along the diagonal line. The vertical line of points at a specific true value signifies that there is an entity with a high frequency and a high degree of prediction error. This entity likely has a large number of members, making it challenging for the model to predict accurately.

On closer examination, this entity turned out to be one marked NIL, which is a catch-all group of all non-linked entities.

The scatter of points away from the diagonal line indicates prediction errors, suggesting that while the model is effective overall, there are certain cases where it does not perform as well.

A. Separate Evalutaion Set

This section presents the evaluation on an independent dataset only for locations. For this purpose, 299 sentences from the novel "The Good Soldier Švejk" by Jaroslav Hašek and 510 sentences from the newspaper "Politika" have been prepared. The evaluation results will be demonstrated by comparing the newly trained model SrpCNNeL with the custom Spacy Entity Linker. We took the strictest approach and differentiated between the following three situations:

- [TP] an entity is recognized exactly as it should, comparing to the gold standard (the text and the QID match – true positives);
- [FP] an entity is recognized, but not with the correct QID;
- [FN] an entity present in the gold standard was not recognized.

The results for the SrpCNNeL model are displayed in the upper part of Table IV.

In the case of the newspaper article, a larger number of entities not linked to a QID but present in the gold standard (FN) are attributed to the entity named Ukraine and all its forms in different grammatical cases, except for the form "Ukrajinu," which is recognized. Upon deeper analysis, it was determined that, at the time of extraction of QIDs, Ukraine did not have the property "instance of state" on Wikidata but instead had other properties such as sovereign state (Q3624078), social state (Q619610), and territory (Q4835091). Consequently, it was not included in the synthetic dataset.

V. Conclusion and Future Work

To the best of our knowledge, this was the first attempt at training a model for the recognition and linking of named entities to a Wikidata knowledge base for the Serbian language, employing spaCy and a CNN network. Although the results on the test set are satisfactory, future research will require expanding the training dataset and applying transformers, which have proven to be more successful than CNN networks in named entity recognition.

One issue observed was the model's difficulty in properly classifying non-linked entities. Increasing the number of linked entities in the knowledge base and expanding the number of properties for extracting QIDs from Wikidata to have more data in the training set would likely improve performance in this area.

Considering that Leximirka has an exhaustive list of locations, a comparison will be made to verify which countries and other categories from Leximirka do not have a corresponding QID. Conversely, an analysis will be conducted to identify which countries, seas, mountains, or organizations exist in the dataset but are missing from Leximirka.

Synthetic sentences have proven to be a valuable source for capturing inflected forms crucial for the morphologically rich Serbian language. Expanding the dataset to include entities with a low frequency of appearance could further enhance the model's performance.

Given the rapid advancements in NLP, CNNs have become somewhat obsolete. Replacing the CNN base layer with BERT or other transformer models is a promising direction for developing a more accurate and robust model.

Acknowledgment

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

TABLE IV
EVALUATION RESULTS ON AN INDEPENDENT DATASET.

ID	TP	FP	FN	P	R	F_1
SrpCNNeL						
novel	29	2	10	0.94	0.74	0.83
newspaper	136	0	84	1.00	0.62	0,76
Baseline Spacy Entity Linker						
novel	2	0	39	1.00	0.05	0.10
newspaper	10	8	202	0.67	0.05	0.09

REFERENCES

[1] K. Balog, *Entity-oriented search*. Springer Nature, 2018. https://doi.org/10.1007/978-3-319-93935-3.

[2] W. Shen, Y. Li, Y. Liu, J. Han, J. Wang, and X. Yuan, "Entity Linking Meets Deep Learning: Techniques and Solutions," 2021. https://doi.org/10.1109/TKDE.2021.3090865.

[3] R. Hanslo, "Evaluation of Neural Network Transformer Models for Named-Entity Recognition on Low-Resourced Languages," in *2021 16th Conference on Computer Science and Intelligence Systems (FedCSIS)*, pp. 115–119, 2021. http://dx.doi.org/10.15439/2021F7.

[4] W. Shen, J. Wang, and J. Han, "Entity linking with a knowledge base: Issues, techniques, and solutions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 2, pp. 443–460, 2014. https://dx.doi.org/10.1109/TKDE.2014.2327028.

[5] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186, Association for Computational Linguistics, 2019. https://doi.org/10.18653/v1/N19-1423.

[6] W. Yin, M. Yu, B. Xiang, B. Zhou, and H. Schütze, "Simple Question Answering by Attentive Convolutional Neural Network," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (Y. Matsumoto and R. Prasad, eds.), (Osaka, Japan), pp. 1746–1756, The COLING 2016 Organizing Committee, 2016. https://doi.org/10.48550/arXiv.1606.03391.

[7] T. Lin, Mausam, and O. Etzioni, "Entity linking at web scale," in *Proceedings of the joint workshop on automatic knowledge base construction and web-scale knowledge extraction (AKBC-WEKEX)*, pp. 84–88, Association for Computational Linguistics, 2012. https://aclanthology.org/W12-3016.

[8] K. Labusch and C. Neudecker, "Named Entity Disambiguation and Linking Historic Newspaper OCR with BERT," in *CLEF (Working Notes)*, p. 33, CEUR-WS, 2020. http://ceur-ws.org/Vol-2696/paper_163.pdf.

[9] Z. Liu, Y. Leng, M. Wang, and C. Lin, "Named Entity Recognition and Named Entity on Esports Contents," in *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, pp. 189–192, 2020. https://doi.org/10.15439/2020F24.

[10] X. Liu, Y. Li, H. Wu, M. Zhou, F. Wei, and Y. Lu, "Entity linking for tweets," in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1304–1311, Association for Computational Linguistics, 2013. https://doi.org/10.1142/9789813227927_0019.

[11] E. French and B. T. McInnes, "An overview of biomedical entity linking throughout the years," *Journal of Biomedical Informatics*, vol. 137, p. 104252, 2023. https://doi.org/10.1016/j.jbi.2022.104252.

[12] R. Sharma, D. Chauhan, and R. Sharma, "Named Entity Recognition System for the Biomedical Domain," in *2022 17th Conference on Computer Science and Intelligence Systems (FedCSIS)*, pp. 837–840, 2022. http://dx.doi.org/10.15439/2022F63.

[13] I. Guellil, A. Garcia-Dominguez, P. R. Lewis, S. Hussain, and G. Smith, "Entity linking for English and other languages: a survey," *Knowledge and Information Systems*, pp. 1–52, 2024. https://doi.org/10.1007/s10115-023-02059-2.

[14] M. Neumann, D. King, I. Beltagy, and W. Ammar, "ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing," in *Proceedings of the 18th BioNLP Workshop and Shared Task*, Association for Computational Linguistics, 2019. https://doi.org/10.18653/v1/w19-5034.

[15] O. Bodenreider, "The unified medical language system (UMLS): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D267–D270, 2004. https://doi.org/10.1093/nar/gkh061.

[16] G. O. Consortium, "The Gene Ontology (GO) database and informatics resource," *Nucleic acids research*, vol. 32, no. suppl_1, pp. D258–D261, 2004. https://doi.org/10.1093/nar/gkh036.

[17] J. M. Van Hulst, F. Hasibi, K. Dercksen, K. Balog, and A. P. de Vries, "Rel: An entity linker standing on the shoulders of giants," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2197–2200, 2020. https://doi.org/10.1145/3397271.3401416.

[18] N. De Cao, L. Wu, K. Popat, M. Artetxe, N. Goyal, M. Plekhanov, L. Zettlemoyer, N. Cancedda, S. Riedel, and F. Petroni, "Multilingual Autoregressive Entity Linking," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 274–290, 2022. https://doi.org/10.1162/tacl_a_00460.

[19] E. Boros, E. L. Pontes, L. A. Cabrera-Diego, A. Hamdi, J. G. Moreno, N. Sidère, and A. Doucet, "Robust named entity recognition and linking on historical multilingual documents," in *Conference and Labs of the Evaluation Forum (CLEF 2020)*, vol. 2696, pp. 1–17, CEUR-WS Working Notes, 2020. https://doi.org/10.5281/zenodo.4068075.

[20] K. Papantoniou, V. Efthymiou, and D. Plexousakis, "Automating Benchmark Generation for Named Entity Recognition and Entity Linking," in *European Semantic Web Conference*, pp. 143–148, Springer, 2023. https://doi.org/10.1007/978-3-031-43458-7_27.

[21] M. Plekhanov, N. Kassner, K. Popat, L. Martin, S. Merello, B. Kozlovskii, F. A. Dreyer, and N. Cancedda, "Multilingual End to End Entity Linking," *arXiv*, 2023. https://doi.org/10.48550/arXiv.2306.08896.

[22] J. Raiman and O. Raiman, "DeepType: Multilingual Entity Linking by Neural Type System Evolution," 2018. https://doi.org/10.48550/arXiv.1802.01021.

[23] P. Nugues, "Linking Named Entities in Diderot's Encyclopédie to Wikidata," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 10610–10615, 2024. https://doi.org/10.48550/arXiv.2406.03221.

[24] N. Loukachevitch, E. Artemova, T. Batura, P. Braslavski, V. Ivanov, S. Manandhar, A. Pugachev, I. Rozhkov, A. Shelmanov, E. Tutubalina, *et al.*, "NEREL: a Russian information extraction dataset with rich annotation for nested entities, relations, and wikidata entity links," *Language Resources and Evaluation*, pp. 1–37, 2023. https://doi.org/10.1007/s10579-023-09674-z.

[25] A. Delpeuch, "Opentapioca: Lightweight entity linking for wikidata," *arXiv preprint arXiv:1904.09131*, 2019. https://doi.org/10.48550/arXiv.1904.09131.

[26] O. Perisic, S. Ranka, I. N. Milica, Š. Mihailo, *et al.*, "It-Sr-NER: CLARIN Compatible NER and GeoparsingWeb Services for Italian and Serbian Parallel Text," in *Selected Papers from the CLARIN Annual Conference 2022, Czechia, 2022*, pp. 99–110, Linköping University Electronic Press, 2023. https://doi.org/10.3384/ecp198010.

[27] O. Perišic, S. Ranka, I. N. Milica, and Š. Mihailo, " It-Sr-NER: Web Services for Recognizing and Linking Named Entities in Text and Displaying Them on a Web Map," *Infotheca - Journal for Digital Humanities*, vol. 23, no. 1, pp. 61–77, 2023. https://doi.org/10.18485/infotheca.2023.23.1.3.

[28] Y. Cao, L. Huang, H. Ji, X. Chen, and J. Li, "Bridge Text and Knowledge by Learning Multi-Prototype Entity Mention Embedding," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1623–1633, Association for Computational Linguistics, 2017. https://doi.org/10.18653/v1/P17-1149.

[29] M. Francis-Landau, G. Durrett, and D. Klein, "Capturing Semantic Similarity for Entity Linking with Convolutional Neural Networks," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (K. Knight, A. Nenkova, and O. Rambow, eds.), (San Diego, California), pp. 1256–1261, Association for Computational Linguistics, 2016. https://doi.org/10.18653/v1/N16-1150.

[30] Y. Shi, R. Yang, C. Yin, Y. Lu, Y. Yang, and Y. Tao, "Entity Linking Method for Chinese Short Texts with Multiple Embedded Representations," *Electronics*, vol. 12, no. 12, 2023. https://doi.org/10.3390/electronics12122692.

[31] R. Pozzi, R. Rubini, C. Bernasconi, and M. Palmonari, "Named Entity Recognition and Linking for Entity Extraction from Italian Civil Judgements," in *International Conference of the Italian Association for Artificial Intelligence*, pp. 187–201, Springer, 2023. https://doi.org/10.1007/978-3-031-47546-7_13.

[32] S. MORAKIS, F. HASIBI, and M. LARSON, "Entity Linking for Greek," 2021.

[33] R. Stanković, C. Krstev, B. Š. Todorović, and M. Škorić, "Annotation of the Serbian ELTeC Collection," *Infotheca–Journal for Digital Humanities*, vol. 21, no. 2, pp. 43–59, 2021. https://doi.org/10.18485/infotheca.2021.21.2.3.

[34] D. Vrandečić and M. Krötzsch, "Wikidata: a free collaborative knowledgebase," *Communications of the ACM*, vol. 57, no. 10, pp. 78–85, 2014. https://doi.org/10.1145/2629489.

[35] B. Lazic and M. Škoric, "From DELA based dictionary to Leximirka lexical database," *Infotheca–Journal for Digital Humanities*, vol. 19, no. 2, pp. 00–00, 2019. https://doi.org/10.18485/infotheca.2019.19.2.4.

[36] D. Hernández, A. Hogan, C. Riveros, C. Rojas, and E. Zerega, "Querying Wikidata: Comparing SPARQL, Relational and Graph Databases," in *The Semantic Web–ISWC 2016: 15th International Semantic Web Conference, Kobe, Japan, October 17–21, 2016, Proceedings, Part II 15*, pp. 88–103, Springer, 2016. https://doi.org/10.1007/978-3-319-46547-0_10.

[37] D. Vitas, S. Koeva, C. Krstev, and I. Obradović, "Tour du monde through the dictionaries," in *Actes du 27eme Colloque International sur le Lexique et la Gammaire*, pp. 249–256, 2008.

[38] C. Krstev, D. Vitas, and A. Trtovac, "Orwells 1984—the Case of Serbian Revisited," in *Proc. of 5th Language & Technology Conference*, pp. 25–27, 2011.

[39] R. Stanković, C. Krstev, D. Vitas, N. Vulović, and O. Kitanović, "Keyword-based search on bilingual digital libraries," in *Semantic Keyword-Based Search on Structured Data Sources: COST Action IC1302 Second International KEYSTONE Conference, IKC 2016, Cluj-Napoca, Romania, September 8–9, 2016, Revised Selected Papers*, pp. 112–123, Springer, 2017. https://doi.org/10.1007/978-3-319-53640-8_10.

[40] M. Ikonić Nešić, S. Petalinkar, S. Ranka, and Š. Mihailo, "BERT downstream task analysis: Named Entity Recognition in Serbian," in *14th International Conference on Information Society and Technology – ICIST 2024*, unpublished, 2024.

[41] M. Škorić, "Novi jezički modeli za srpski jezik," *Infotheca - Journal for Digital Humanities*, 2024. https://doi.org/10.48550/arXiv.2402.14379.

[42] Y. Liu, M. Ott, N. Goyal, *et al.*, "Roberta: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019. https://doi.org/10.48550/arXiv.1907.11692.

[43] J.-C. Klie, M. Bugert, B. Boullosa, *et al.*, "The INCEPTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation," in *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pp. 5–9, 2018.

[44] B. Šandrih Todorović, C. Krstev, R. Stanković, and M. Ikonić Nešić, "Serbian NER& Beyond: The Archaic and the Modern Intertwinned," in *Deep Learning Natural Language Processing Methods and Applications – Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2021)* (G. Angelova, M. Kunilovskaya, R. Mitkov, and I. Nikolova-Koleva, eds.), pp. 1252–1260, INCOMA Ltd., September 2021. https://doi.org/10.26615/978-954-452-072-4_141.

[45] R. Stanković, C. Krstev, B. Šandrih Todorović, and M. Škorić, "Annotation of the Serbian ELTeC Collection," *Infotheca - Journal for Digital Humanities*, vol. 21, no. 2, pp. 43–59, 2021. https://doi.org/10.18485/infotheca.2021.21.2.3.

[46] B. Lazić and M. Škorić, " From DELA based dictionary to Leximirka lexical database," *Infotheca - Journal for Digital Humanities*, vol. 19, no. 2, pp. 81–98, 2020. https://10.18485/infotheca.2019.19.2.4.