Serbian NER&Beyond: The Archaic and the Modern Intertwinned

Branislava Šandrih Todorović, Cvetana Krstev, Ranka Stanković, Milica Ikonić Nešić

Serbian NER&Beyond: The Archaic and the Modern Intertwinned

Branislava Šandrih Todorović
University of Belgrade
Faculty of Philology
branislava.sandrih@fil.bg.ac.rs

Cvetana Krstev
University of Belgrade
Faculty of Philology
cvetana@matf.bg.ac.rs

Ranka Stanković
University of Belgrade
Faculty of Mining and Geology
ranka.stankovic@rgf.bg.ac.rs

Milica Ikonić Nešić
University of Belgrade
Faculty of Philology
milica.ikonic.nesic@fil.bg.ac.rs

Abstract

In this work, we present a Serbian literary corpus that is being developed under the umbrella of the "Distant Reading for European Literary History" COST Action CA16204. Using this corpus of novels written more than a century ago, we have developed and made publicly available a Named Entity Recognizer (NER) trained to recognize 7 different named entity types, with a Convolutional Neural Network (CNN) architecture, having F_1 score of ≈91% on the test dataset. This model has been further assessed on a separate evaluation dataset. We wrap up with comparison of the developed model with the existing one, followed by a discussion of pros and cons of the both models.

1 Introduction

The "Distant Reading for European Literary History"[1] (COST Action CA16204) has started in 2017 with the purpose of using computational methods to analyse large collections of literary texts (Stanković et al., 2019; Frontini et al., 2020). The main goal of this ongoing action is to compile a multilingual open-source collection, named European Literary Text Collection (ELTeC), containing linguistically annotated sub-collections of 100 novels per language written more than 100 years ago.

In this paper, we present a collection of Serbian texts in this corpus, named SrpELTeC. Alongside, we describe our efforts in developing its Named Entity (NE) layer, defined previously as one of the main action's deliverables.

For this purpose, we adjusted and used the existing rule-based NE recognizer for Serbian, dubbed SrpNER, that we will describe in Section 2 together with some approaches to NE recognition in literary texts. This SrpNER model was applied to the raw version of the selected texts from SrpELTeC collection, presented in Section 3. Based on the specifically tailored guidelines, different evaluators performed careful checks and corrections, yielding a gold standard (SrpELTeC-gold). This enabled us to train a CNN-based NE recognizer, named SrpCNNER, presented in Section 4. Having the gold dataset, prepared as described in Subsection 4.1, we trained (Subsection 4.2) and evaluated the model in two different settings: first, we discussed our model's performance on the SrpELTeC-gold test subset, as shown in Subsection 4.3. Afterwards we carried out a detailed evaluation on a collection of novels that were not present in the gold standard, named SrpELTeC-eval, with the findings and a thorough discussion given in Section 5. Finally, conclusions and plans for the future work were stated in Section 6.

2 Related Work

The existence of large-scale lexical resources for Serbian, e-dictionaries in particular (Krstev, 2008), coupled with local grammars in the form of finite-state transducers (Vitas and Krstev, 2012), enabled the development of a comprehensive rule-based system for NER SrpNER. This system presented by Krstev et al. (2014) targeted 11 classes of NEs: dates and time (moments and periods), money and measurement expressions, geopolitical names (countries, settlements, oronyms and hydronyms), and personal names (one or more last names with or without first names and nicknames). The system was developed to recognize NEs in

[1] Distant Reading, https://www.distant-reading.net

Proceedings of Recent Advances in Natural Language Processing, pages 1252–1260
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_141

newspapers and similar texts. It was manually evaluated on a sample of unseen newspaper texts. The overall F_1 score of the model was $\approx 96\%$. To the best of our knowledge, so far there were no attempts to produce a NER system for Serbian literary texts.

The enhanced version of SrpNER was later utilized by Šandrih et al. (2019) for the preparation of a gold standard annotated with personal names, which was used for building training sets for 4 different levels of annotation, on which two ML-based NE recognizers were trained and evaluated (SpaCy and Stanford). As a support for the developed NER models, Šandrih et al. (2019) joined several existing tools and developed various new tools, combined into a web platform NER&Beyond.[2]

Although NER systems in general were developed mostly for newspaper and similar texts, there were some endeavours to produce functional systems for literary texts as well. Enrichment of French Renaissance texts with proper names (Maurel et al., 2014) faced two challenges: text diversity due to various spellings of words, and need to deal with numerous XML-TEI tags used to preserve the format of original editions. Authors' solution was based on the cascades of finite-state automata and both general dictionaries and those built specifically for the project. The evaluation showed that the slot error rate of name tagging was 6.1%.

A dataset of literary entities comprising 210,532 tokens evenly drawn from 100 different English literary texts annotated with ACE entity categories (person, location, geo-political entity, facility, organization, and vehicle)[3] was published in (Bamman et al., 2019). The authors' main motivation was to asses NER models' performance on different types of texts. Their conclusion was that recognition improved for almost all entity types when literary texts were used for the both training and evaluation (on average $P = 75.1\%$, $R = 62.6\%$ and $F_1 = 68.3\%$), whilst for training on general texts, such as news data, and testing on literary texts the results were much poorer (on average $P = 57.8\%$, $R = 37.7\%$ and $F_1 = 45.7\%$).

SHINRA2020-ML shared-task (Sekine et al., 2020) targeted the categorization of Wikipedia entities using the Extended Named Entity (ENE) hierarchy in 30 languages (Serbian was not one of them). ENE included about 220 fine-grained categories of NEs in a hierarchy of up to four layers. Some traditional NE types such as location were specified as either geopolitical location ("city", "province", "country", etc.) or geological region ("mountain", "river", "lake", etc.). ENE also included some new NE types like "products", "event", "position", etc.

Dekker et al. (2019) experimented with different off-the-shelf NER tools for the extraction of social network graphs from classic and modern English fiction novels. The authors wanted to find out to what extent are these tools suitable for identifying fictional characters in novels, and what are differences and similarities that can be discovered between social networks extracted for different novels.

Distant Reading Training School for Named Entity Recognition and Geo-Tagging for Literary Analysis organized within the COST Action 16204[4] covered NER approaches in general, annotation campaigns, practical work with NER tools, annotating NER in TEI, analyzing NER annotation for literary characters and place names and NER data analysis. Different types of NER systems were tested for several languages, some based on symbolic methods, relying on rules developed by experts and dictionaries (gazetteers), others using statistical and data-driven approach.

The NE layer of ELTeC corpus has presently been produced for three languages: Hungarian, Portuguese, Slovene. Santos et al. (2020) reported on the NER annotation of the Portuguese sub-collection of the ELTeC corpus. Authors used the PALAVRAS-NER parser, a Constraint Grammar (CG) system, in which NER is an integrated task of grammatical tagging, implemented with the basic tagset of 6 NE categories (person, organization, place, event, semantic products and objects) with about 20 subcategories at three levels, disambiguated by CG-rules: known lexical entries and gazetteer lists, pattern-based name type prediction and context-based name type inference for unkno-

[2] NER&Beyond, http://nerbeyond.jerteh.rs/

[3] ACE (Automatic Content Extraction) 2005 Multilingual Training Corpus, https://catalog.ldc.upenn.edu/LDC2006T06

[4] Materials for the NER Training School, https://github.com/distantreading/WG2/tree/master/NER_TS

wn words. This system was applied to eight novels that were fully human revised. Evaluation results varied for precision from 64.6% to 80.8%, and recall from 64.3% to 82.0%.

At the mentioned Distant reading training school it was concluded that spaCy module[5] for Python was used for training NER models for many involved languages, already having tagsets that could be mapped to the ELTeC annotation scheme, elaborated later in Section 3. Partalidou et al. (2019) developed a POS-tagger and a NER for Greek using spaCy, based on newspaper articles and Wikipedia dataset, able to recognize the following entity types: location, organization, person and facility. Jabbari et al. (2020) created a corpus consisting of news articles in French, which served as a dataset for training and evaluation of a NER and a relation extraction algorithms using spaCy. Modrzejewski et al. (2020) incorporated NER trained in spaCy into an English/German Machine Translation system, with the aim to improve NE translation.

Morever, Jiang et al. (2016) conducted a comparative evaluation of different publicly available NER tools. Based on different criteria, authors concluded that spaCy was among best performing across all tested datasets. Having all this in mind, we decided for spaCy as a framework for developing a Serbian NER model on a collection old literary texts.

3 Serbian Collection in the ELTeC

As described earlier in Section 1, the focus of the COST Action CA16204 is to compile the ELTeC corpus containing collections of old European novels published between 1840 and 1920 in various languages. In order to make these sub-collections decent representatives of their corresponding languages, the novels were selected to evenly represent a) novels of various sizes: short, medium, long; b) four twenty-year time periods within the examined time span, c) canonical novels as well as those not known to wider audience or completely forgotten, as judged by the number of reprints, and d) female and male authors (Frontini et al., 2020).

The last version of the ELTeC (v. 1.1.0) was released in April 2021.[6] It contained 14 language sub-collections each with at least 50 novels, while 8 collections contained targeted 100 novels per language.

The SrpELTeC corpus[7] in the latest ELTeC release has 90 novels. The work on this collection is still in progress with the aim to obtain the complete collection by the end of the project. Contrary to a number of other European languages involved in this action, the Serbian corpus is being produced from scratch, because the vast majority of novels from the selected time period were not digitized before, they were not digitized in the proper manner or were not available (Krstev et al., 2019).

This preparation procedure involved several steps: selection of novels, retrieval of hard copies, scanning, OCR, automatic correction of OCR errors (for which a specialized tool based on the Serbian morphological dictionaries was produced (Krstev and Stanković, 2020)), correction of remaining errors by a number of volunteer readers, and production of metadata.

One of the important aspects of this ELTeC collection is to feature annotations of certain named entities. At this moment, annotation of named entities is carried out for nine languages, including Serbian. According to the guidelines, the common NER tagset includes the following 7 categories: demonyms (DEMO), professions and titles (ROLE), works of art (WORK), person names (PERS), places (LOC), events (EVENT) and organizations (ORG).[8]

4 SrpCNNER Model for Serbian

In this section we first explain how we have turned the SrpELTeC corpus into a dataset for NER. Afterwards, we describe the training of the NER model SrpCNNER, followed by a detailed evaluation. Web users can navigate to http://ner.jerteh.rs/ in order to apply the SrpCNNER model directly on input text. The model can also be applied to a custom-size collection of text files using the previously mentioned NER&Beyond web platform.

[5] spaCy, https://spacy.io/

[6] ELTeC (Distant Reading for European Literary History), https://zenodo.org/communities/eltec

[7] SrpELTeC, https://distantreading.github.io/ELTeC/srp/index.html

[8] ELTeC Collections with NE-annotations, http://brat.jerteh.rs/index.xhtml#/eltec/

4.1 Gold Standard: SrpELTeC-gold

The SrpNER system for Serbian introduced in Section 2 was used in the first stage of the gold standard preparation (dubbed SrpELTeC-gold) in order to automatically annotate SrpELTeC collection. The tagset used by SrpNER differed from the simplified tagset used in the ELTeC project – the tags are more refined, e.g. toponyms are classified as oronyms, hydronyms, settlements etc., and nesting of tags is allowed. Thus, the tags produced by SrpNER had to be mapped to ElTeC tags as illustrated in Figure 1:

```
SrpNER:
<pers.spec>
 <role>
    <demonym>ruskog</demonym> cara
 </role>
 <persName.first>Nikolu</persName.first>
</pers.spec>
ELTeC:
<DEMO>ruskog</DEMO>
<ROLE>cara</ROLE>
<PERS>Nikolu</PERS>
```

Figure 1: SrpNER tags mapped to ELTeC tags (Russian tzar Nikolai).

Before text annotation, we used the advantage of rule-based NER systems and adjusted SrpNER to these specific texts that differ significantly from newspaper texts for which SrpNER was primarily developed in order to improve its performance and facilitate the work of evaluators. Some modifications of rules and used lexicons were done for the whole collection (e.g. *Danas* 'today' cannot be the name of an organization since this publishing house was established 20 years ago), while others were novel-specific (e.g. *Una* can be the first name or the name of the river – we retained only the possibility appropriate to the particular novel).

The EVENT named entity is somewhat special: SrpNER does not recognize this entity, so the evaluators were asked to identify and annotate them when they occur in text. SrpNER does not recognize WORK entity either, but these annotations were in many cases added by volunteer readers during text correction.

Afterwards, students were given different novel chapters along with the annotation guidelines presented briefly in Table 1. Following these instructions and under constant supervision of their professors, students manually corrected the automatically annotated chapters.

The evaluators were divided into two groups: the first group performed corrections using the BRAT annotation tool,[9] while the second group used the INCEpTION.[10] We wanted to receive user feedback on both platforms for the sake of creating the annotation process as comfortable and efficient as possible in the future, but also to provide choice to annotators. The fundamental difference was the input format these platforms needed: BRAT tool uses the standoff format, whilst INCEpTION relies on the CoNLL-2002 verticalized format.[11] In order to convert from one format into another, we used the NER&Beyond web application.

Table 2 displays distribution of different entity types over SrpELTeC-gold novels. The first four digits of text identifiers represent the year of the first publication of a novel. For some novels, NER was not performed on the whole text, but rather on randomly selected chapters. These annotated samples were also included in the gold standard. The cumulative values of entities on all samples are indicated in the first row (ID "sample"). Column $\sum_{tok}$ indicates a novel's size in terms of tokens.

4.2 Training

We trained our SrpCNNER model on the SrpELTeC-gold corpus using the spaCy Python module, version 3.0. In order to prepare the dataset for training, we first segmented texts into sentences, ending up with 43,129 sentences in total, including sentences that did not contain named entities. Afterwards, we randomly shuffled and split these sentences into training, test and development sets with the ratio of 8:1:1, i.e. 34,503 sentences in the train set, and the same number of sentences, 4,313, in the test and development sets, respectively.

These sentences were prepared as Python list-objects containing tuples as elements. An example of such tuple is the following:

"Hadži-Đera je za to vreme ušao u sobu agama, da im nazove dobro jutro, a manastir-

[9] BRAT, https://brat.nlplab.org

[10] INCEpTION annotation tool, https://inception-project.github.io/

[11] Among other CoNLL and XML variants that this tool supports.

Entity	Explanation
PERS Personal names	First names, surnames, nicknames and their combinations (of real people and fictional characters, including gods and saints). Possessive adjectives from personal names should not be annotated.
ROLE Occupations and titles	Occupations, titles and responsibilities: doctor, teacher; king; director.
LOC Locations	Continents, countries, regions, populated places, oronyms, water surfaces, names of celestial bodies, city locations.
DEMO Origin or residence	Residents of states, cities, regions, or ethnic groups; adjectives derived from the names of locations.
ORG Organizations, institutions, societies	Company names, politic parties, educational institutions, sport teams, hospitals, museums, libraries, hotels, cafes, churches and shrines.
WORK Art works	Titles of books, plays, poems, paintings, sculptures, newspapers.
EVENT Events	Names of events that are repeated regularly or have happened once but have their own name: natural disasters, revolutions, battles, wars.

Table 1: Annotation guidelines.

ID	PERS	ROLE	LOC	DEMO	ORG	WORK	EVENT	$\sum_{tok}$
samples	707	207	156	105	8	4	14	19,274
18750	1,688	1,050	388	239	29	10	21	31,743
18871	1,612	1,509	328	229	52	60	18	34,324
18880	1,372	986	271	201	32	59	10	26,642
18881	935	619	95	105	12	14	1	13,898
18890	804	714	36	56	1	0	0	29,337
18932	1,521	259	46	35	0	5	2	16,821
18950	764	581	51	103	12	6	33	14,454
19021	1,647	2,285	123	58	82	4	15	40,804
19040	1,655	917	221	281	1	3	7	32,367
19140	770	412	240	94	45	5	7	31,583
19190	1,181	797	8	13	49	24	19	33,562
total	14,788	10,405	1,979	1,568	323	198	149	330,119

Table 2: SrpELTeC-gold NE distribution.

ski sluga poče prisluživati rakiju i kafu.",[12] 'entities': [(0, 10, 'PERS'), (39, 44, 'ROLE'), (86, 91, 'ROLE')]

The spaCy v3.0 enables specification of custom neural network architecture within a simple text file. Using the quick-start widget,[13] user can easily set up the default setting configuration. In our case, the model's language was Serbian, containing the *ner* component only, trained on CPU. We made the following adjustments to the default configuration (referring to the corresponding file blocks):

[components.tok2vec.model.encode] changed size of the token-to-vector layer from 96 to 300, that is maximum recommended value (`width` parameter);

[components.ner.model] changed width of a hidden layer from 64 to to 300 (`hidden_width` parameter);

[12] Translates as: *In the meantime, Haji-Đera entered the room to wish agas good morning, when the monastery servant started offering coffee and brandy.*

[13] Quick-start spaCy3 widget, https://spacy.io/usage/training#quickstart

[components.ner.model.tok2vec] set the architecture (`@architectures`) to `HashEmbedCNN`[14] having width of the input and the output equal to 300 (`width`), with 8 convolutional layers (`depth`), 10,000 rows in the hash embedding tables (`embed_size`), with the recommended 1 token on either side to concatenate during the convolutions (`window_size`), without pretrained static vectors (`pretrained_vectors = null`).

Model training ended up after 11 epochs (the number of epochs is automatically generated), having 93.33%, 90.14% and 91.71% F_1 score, precision and recall, on the development set, respectively.

4.3 Evaluation

Afterwards, we examined our model's performance on the test set. We run the previously trained model on raw, non-annotated sentences from the SrpELTeC. After comparing the obtained annotations with the ones given in the test subset of the SrpELTeC-gold, we obtained the precision (P), recall (R) and F_1 scores displayed in Table 3.

Type	P	R	F_1
PERS	0.953	0.936	0.944
ROLE	0.940	0.917	0.928
LOC	0.849	0.778	0.812
DEMO	0.781	0.758	0.769
ORG	0.903	0.368	0.523
WORK	0.324	0.343	0.333
EVENT	0.792	0.655	0.717

Table 3: SrpCNNER on the test set.

The normalized confusion matrix is given in Figure 2 ('O' represents tokens that are not NE). One can observe that WORK and EVENT were frequently missed or confused with PERS.

5 Separate Evaluation Set

Despite the encouraging results obtained on the SrpELTeC-gold, shown in Subsection 4.3, we

[14] HashEmbedCNN, https://spacy.io/api/architectures#HashEmbedCNN

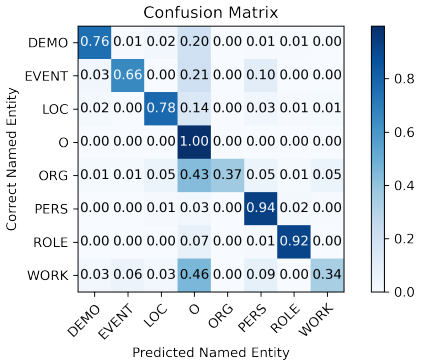

Figure 2: Confusion matrix on the test set.

wanted to further asses our model's performance. For this purpose, we prepared an independent evaluation set, dubbed SrpELTeC-eval, containing corrected annotated chapters from three novels that were not included in the training procedure. Table 4 displays entity distribution over SrpELTeC-eval. Named entities are represented by their first letter (e.g. P represents PERS). It should be noted that the EVENT type did not occur in this dataset.

ID	P	R	L	D	O	W	$\sum_{tok}$
19070	44	55	23	23	3	0	2,027
19180	18	13	2	5	0	5	3,928
19121	33	18	14	2	0	0	3,045

Table 4: SrpELTeC-eval NE distribution.

We applied the same evaluation procedure for the both recognizers. After running a them on SrpELTeC-eval, 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 named entity types match – true positives);

[FP] there are three cases here: 1) an entity is recognized, but not with the correct type (e.g. PERS mistaken for a ROLE); 2) an entity is recognized as a correct type but the scope is not correct (e.g only a first name is recognized as PERS, although a full name is given); or 3) model annotated something that is not present in the gold standard – false positives;

[FN] an entity present in the gold standard was not recognized – false negatives.

In the subsections that follow, we analyze the performances of our newly trained model SrpCNNER and the adjusted SrpNER on the SrpELTeC-eval corpus. Finally, we discuss their strengths and weaknesses and make certain statements about their applicability in different contexts and situations.

5.1 SrpCNNER vs. SrpELTeC-eval

The overall results for the SrpCNNER are displayed in the upper part of Table 5. As previously explained, for the case of FP, there is a specific situation that something was recognized, but not with the correct entity type. Such cases are indicated by the number in parentheses of the FP column (therefore, numbers TP, FN and the one given in parentheses from the FP column sum up to the total number of entities given in the $\sum$ column in Table 4).

ID	TP	FP	FN	P	R	F_1
SrpCNNER vs. SrpELTeC-eval						
19070	50	25(18)	80	0.538	0.385	0.448
19180	27	29(4)	12	0.450	0.692	0.545
19121	34	23(6)	27	0.540	0.557	0.548
SrpNER vs. SrpELTeC-eval						
19070	128	5(2)	18	0.948	0.877	0.911
19180	27	24(2)	14	0.509	0.659	0.574
19121	47	15(0)	20	0.758	0.701	0.729

Table 5: Evaluation results SrpELTeC-eval.

Values of precision (P), recall (R) and F_1 scores over each entity are shown in the upper part of Figure 3.

5.2 SrpNER vs. SrpELTeC-eval

The overall results for the SrpNER are displayed in the lower part of Table 5. Values of precision (P), recall (R) and F_1 scores over each entity are shown in the lower part of Figure 3.

From the obtained results it is obvious that SrpNER was not nearly as successful as when applied to newspaper texts. This could well be expected since each novel has its own specifics, and one cannot say that novels in general share some common language features, as newspapers do. Also, one can observe that results are very different for each of three samples; however, we cannot draw some firm conclusions, since the used samples were rather small.

5.3 Discussion

Based on the results shown in Figure 3 (upper part) and Table 5, it becomes obvious that SrpCNNER does not perform so well on unseen texts. In order to understand the reasons for that, we observed each and single case in isolation, which brought us to certain findings.

SrpCNNER performed rather well in recognizing personal names (e.g. *Ana*, *Nikola*, *Gavra Đaković*, *Ismail*), roles and titles (e.g. *car* 'tsar', *sultan*, *princeza* 'princess', *sveštenik* 'priest'), locations (e.g. *Beograd*, *Pariz*, *Niš*), and demonyms (e.g. *Švaba* 'German' (pejorative), *ruskom* 'Russian', *francuskom* 'French'). However, the number of FP cases was intriguing, due to the ambiguity of use. For example, the model recognized all occurrences of the word *otac* 'father' as a ROLE, although it can represent both a male parent (which according to the guidelines should not be annotated) and a priest (which should be annotated). Similar is the case with *čika* 'uncle', which in Serbian, when used before a personal name, has the meaning of mister/sir (familiarly). Both words are used rather frequently, and out of 33 false positives for the novel 19180, 13 were occurrences of exactly these two words.

The novel 19070 revealed some new weak points. For example, occurrences such as *Fati-Sultan*, *Ismail-beg* and *Ahmed-hafuz* are specific to this novel and they represent a combination of a PERS-ROLE entities, a construction that is not usual in Serbian – ROLE PERS order is preferred. SrpCNNER recognized these two entities as a single PERS, WORK or LOC entity (among 43 false positives for the 19070, 7 were these names in various inflected forms), or did not recognize them at all (14 times).

We also noticed that some false positives were due to specific characteristics of texts. Namely, the orthography in the old novels was not stable, leading to incorrect occurrences (according to contemporary usage); for instance, the word *gospode* 'god' was considered, according to the decision of the evaluator, FP because written with the lower-case G, while the same word written with the upper-case G *Gospode* 'God' was found among the true positives.

It should be noted that in literary texts it is not always easy to decide what is the right type of an NE. For instance, in a sentence from

Figure 3: SrpCNNER vs. SrpELTeC-eval (upper) and SrpNER vs. SrpELTeC-eval (lower).

19180: *Sa Tolstojem sam se pomirila i obožavam ga za Anu Karenjinu* 'I reconciled with Tolstoy and I adore him for Anna Karenina', *Ana Karenjina* can refer to the novel (WORK) or to its main character (PERS), and it is open to interpretation. Similarly, the names of saints (PERS) were sometimes difficult to distinguish from festivities that celebrate them (EVENT). One such example from 18950 is: *Mi slavimo Svetog Nikolu, ovog letnjeg.* 'We celebrate Saint Nicolas, the one that comes in summer.'

Finally, we have noticed that our gold standard has flaws, introduced by evaluators, especially when facing some of the tricky cases mentioned before. It would have certainly been better if we could engage two evaluators for each text, but our human resources were limited.

Overall conclusion is that SrpCNNER performs satisfactorily on similar texts, which can be seen from the model's performance on the test set displayed in Table 3. Since this collection of novels contains very diverse texts, both lexically and syntactically, SrpCNNER did not generalize that well on unseen texts.

6 Conclusions and Future Work

We presented the corpus of old Serbian novels, which served as a basis for training a CNN-based NER model SrpCNNER using the spaCy module's framework for Python. After comparing this newly developed model for Serbian with the existing rule-based SrpNER, we came to the conclusion that the previously developed one performs better on this type of texts, due to its adaptability. However, it is not easy to set it up and use it, while the model trained in spaCy can be easily and efficiently applied to the large text collections, and there is still a lot of room for improvement. First of all we need to remove observed flaws from SrpELTeC-gold. Moreover, in the future we intend to use the pre-trained word embedding vectors instead of the default `tok2vec` layer.

The integration of POS-tagging and lemmatization with NER into TEI ELTeC level 2 schema[15] is an ongoing activity, where a pipeline starts with SrpNER annotation, followed by POS-tagging and lemmatization by a TreeTagger (Schmid, 1999; Stanković et al., 2020). As a result, first 16 novels from SrpELTeC collection were annotated with POS, lemmas, and NE in a format agreed by the COST action.

Acknowledgments

This research was done in the scope of the COST action CA16204 "Distant Reading for European Literary History". We thank the students of the Department of Library and Information Sciences, Faculty of Philology, master students of Social Sciences and Computing at Multidisciplinary Graduate Studies and PhD students of Intelligent systems program (University of Belgrade) for their help in evaluating the data.

[15] Encoding Guidelines for the ELTeC: level 2, https://distantreading.github.io/Schema/eltec--2.html

References

David Bamman, Sejal Popat, and Sheng Shen. 2019. An Annotated Dataset of Literary Entities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 2138–2144.

Niels Dekker, Tobias Kuhn, and Marieke van Erp. 2019. Evaluating Named Entity Recognition Tools for Extracting Social Networks from Novels. *PeerJ Computer Science*, 5:e189.

Francesca Frontini, Carmen Brando, Joanna Byszuk, Ioana Galleron, Diana Santos, and Ranka Stanković. 2020. Named Entity Recognition for Distant Reading in ELTeC. In *CLARIN Annual Conference 2020*.

Ali Jabbari, Olivier Sauvage, Hamada Zeine, and Hamza Chergui. 2020. A French Corpus and Annotation Schema for Named Entity Recognition and Relation Extraction of Financial News. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2293–2299, Marseille, France. European Language Resources Association.

Ridong Jiang, Rafael E Banchs, and Haizhou Li. 2016. Evaluating and Combining Name Entity Recognition Systems. In *Proceedings of the 6th Named Entity Workshop*, pages 21–27.

Cvetana Krstev. 2008. *Processing of Serbian. Automata, Texts and Electronic Dictionaries*. Faculty of Philology of the University of Belgrade.

Cvetana Krstev, Jelena Jaćimović, Branislava Šandrih, and Ranka Stanković. 2019. Analysis of the first Serbian Literature Corpus of the Late 19th and Early 20th century with the TXM platform. In *DH_BUDAPEST_2019*, pages 36–37. Centre for Digital Humanities - Eötvös Loránd University. http://elte-dh.hu/wp-content/uploads/2019/09/DH_BP_2019-Abstract-Booklet.pdf.

Cvetana Krstev, Ivan Obradović, Miloš Utvić, and Duško Vitas. 2014. A System for Named Entity Recognition Based on Local Grammars. *Journal of Logic and Computation*, 24(2):473–489.

Cvetana Krstev and Ranka Stanković. 2020. Old or New, we Repair, Adjust and Alter (Texts). *Infotheca - Journal for Digital Humanities*, 19(2):61–80.

Denis Maurel, Nathalie Friburger, and Iris Eshkol-Taravella. 2014. Enrichment of Renaissance Texts with Proper Names. *INFOtheca: Journal of Information and Library Science*, 15(1):15–27.

Maciej Modrzejewski, Miriam Exel, Bianka Buschbeck, Thanh-Le Ha, and Alexander Waibel. 2020. Incorporating External Annotation to improve Named Entity Translation in NMT. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 45–51, Lisboa, Portugal. European Association for Machine Translation.

Eleni Partalidou, Eleftherios Spyromitros-Xioufis, Stavros Doropoulos, Stavros Vologiannidis, and Konstantinos Diamantaras. 2019. Design and Implementation of an Open Source Greek POS–Tagger and Entity Recognizer Using spaCy. In *IEEE/WIC/ACM International Conference on Web Intelligence*, WI '19, page 337–341, New York, NY, USA. Association for Computing Machinery.

Diana Santos, Eckhard Bick, and Marcin Wlodek. 2020. Avaliando Entidades Mencionadas na Coleção ELTeC-por. *Linguamática*, 12(2):29–49.

Helmut Schmid. 1999. Improvements in Part-of-Speech Tagging with an Application to German. In *Natural language processing using very large corpora*, pages 13–25. Springer.

Satoshi Sekine, Masako Nomoto, Kouta Nakayama, Asuka Sumida, Koji Matsuda, and Maya Ando. 2020. Overview of SHINRA2020-ML Task. In *Proceedings of the NTCIR-15 Conference*.

Ranka Stanković, Branislava Šandrih, Cvetana Krstev, Miloš Utvić, and Mihailo Škorić. 2020. Machine Learning and Deep Neural Network-Based Lemmatization and Morphosyntactic Tagging for Serbian. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3954–3962.

Ranka Stanković, Diana Santos, Francesca Frontini, Tomaz Erjavec, and Carmen Brando. 2019. Named Entity Recognition for Distant Reading in Several European Literatures. In *DH Budapest 2019*.

Duško Vitas and Cvetana Krstev. 2012. Processing of Corpora of Serbian Using Electronic Dictionaries. *Prace Filologiczne*, 63:279–292.

Branislava Šandrih, Cvetana Krstev, and Ranka Stanković. 2019. Development and Evaluation of Three Named Entity Recognition Systems for Serbian - The Case of Personal Names. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1060–1068, Varna, Bulgaria. INCOMA Ltd.