

Development and Evaluation of Three Named Entity Recognition Systems for Serbian - The Case of Personal Names

Branislava Šandrih, Cvetana Krstev, Ranka Stanković



Дигитални репозиторијум Рударско-геолошког факултета Универзитета у Београду

[ДР РГФ]

Development and Evaluation of Three Named Entity Recognition Systems for Serbian - The Case of Personal Names |
Branislava Šandrih, Cvetana Krstev, Ranka Stanković | Proceedings - Natural Language Processing in a Deep Learning
World | 2019 | |

10.26615/978-954-452-056-4_122

<http://dr.rgf.bg.ac.rs/s/repo/item/0005243>

Дигитални репозиторијум Рударско-геолошког факултета Универзитета у Београду омогућава приступ издањима Факултета и радovima запослених доступним у слободном приступу. - Претрага репозиторијума доступна је на www.dr.rgf.bg.ac.rs

The Digital repository of The University of Belgrade Faculty of Mining and Geology archives faculty publications available in open access, as well as the employees' publications. - The Repository is available at: www.dr.rgf.bg.ac.rs

Development and Evaluation of Three Named Entity Recognition Systems for Serbian - The Case of Personal Names

Branislava Šandrih

University of Belgrade
Faculty of Philology
Belgrade, Serbia

branislava.sandrih@fil.bg.ac.rs

Cvetana Krstev

University of Belgrade
Faculty of Philology
Belgrade, Serbia

cvetana@matf.bg.ac.rs

Ranka Stanković

University of Belgrade
Faculty of Mining and Geology
Belgrade, Serbia

ranka@rgf.bg.ac.rs

Abstract

In this paper we present a rule- and lexicon-based system for the recognition of Named Entities (NE) in Serbian newspaper texts that was used to prepare a gold standard annotated with personal names. It was further used to prepare training sets for four different levels of annotation, which were further used to train two Named Entity Recognition (NER) systems: Stanford and spaCy. All obtained models, together with a rule- and lexicon-based system were evaluated on two sample texts: a part of the gold standard and an independent newspaper text of approximately the same size. The results show that rule- and lexicon-based system outperforms trained models in all four scenarios (measured by F_1), while Stanford models have the highest recall. The produced models are incorporated into a Web platform NER&Beyond that provides various NE-related functions.

1 Introduction

Named Entity Recognition is the task of identifying named entities in text (Nadeau and Sekine, 2007), which is often used as a first step in question answering, information retrieval, anaphora resolution, topic modeling, etc. The first Named Entity set had 7 types (Grishman and Sundheim, 1996): organization, location, person, date, time, money and percent expressions. Sekine et al. (2002) proposed a NE hierarchy which contains about 150 NE types.

There are three categories of NER systems:

1) The rule-based (RB) (Krupka and Hausman, 1998; Friburger and Maurel, 2004); 2) the Machine Learning (ML) based (Finkel and Manning,

2009; Singh et al., 2010); and 3) hybrid methods (Jansche and Abney, 2002). The ML-based methods can often be “black boxes”, in comparison with RB techniques which are easy to interpret. Yet, ML-based methods are state-of-the-art. Such example is a Stanford Named Entity Recognizer (Manning et al., 2014), which can be trained for many languages. Other notable NER platforms include GATE (Desktop application that enables NER across many languages and domains),¹ OpenNLP (rule-based and statistical NER),² spaCy (Honnibal and Montani, 2017) (module written in Python, used for advanced NLP)³ and many others.

For Serbian, thus far a rule-based and lexicon-based NER system was developed – SRPNER (Krstev et al., 2014). Its development started with the recognition of a NE class present in all NE schemes, personal names (Krstev et al., 2005), while the recognition of other main NE classes was subsequently added. In the next Section we present briefly this system and how it was used to produce the corpus of newspaper texts annotated with personal names – the gold standard. Section 3 describes NER systems based on Machine Learning methods that were trained on the corpus derived from the gold standard, while the evaluation and discussion of results are presented in Sections 4 and 5, respectively. In Section 6 we present a web platform that enables use and evaluation of these systems. Finally, some directions for future work are given in the last Section.

¹GATE, <https://gate.ac.uk/>

²OpenNLP, <https://opennlp.apache.org/>

³spaCy, <https://spacy.io/>

2 SRPNER and the Gold Standard

2.1 Rule-Based NER for Serbian

The first NER system for Serbian was a rule- and lexicon-based system developed several years ago. It has been designed to recognize the main classes of NEs: 1) numerical expressions (measurement and money), 2) temporal expressions (date and time, and 3) name expressions (personal, geopolitical and organization names).

The system was designed in a form of the cascades of Finite-State Transducers (FST) in which every transducer recognizes and tags a certain class of NEs (Friburger and Maurel, 2004; Maurel et al., 2011). Each transducer rely in its work on the results of previous transducers and on e-dictionaries of Serbian (Vitas and Krstev, 2012). E-dictionaries play an important role specifically in the recognition of name expressions, since, beside general lexica, they contain many proper names, both personal and geopolitical. The system is modular which means that steps can be omitted and can change order; however, it performs best when used in predefined way since in each step the disambiguation of some names is performed.

SRPNER presented in (Krstev et al., 2014) recognizes 11 classes of NEs: dates (moments and periods), time (moments and periods), money expressions, 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 presented evaluation results for the recognition of all mentioned NEs obtained on a sample of newspaper texts were $F_1 = 0.96$ ($R = 0.94$; $P = 0.98$). The system also recognized titles, roles and functions of persons when they accompany personal names.

Since this first results, system has been continually improved by adding new NE classes (organization names) and new sub-classes (e.g. for geopolitical names: regions, super-regions and city counties). In addition, the e-dictionaries of Serbian were also continually improved and enhanced, and that by itself contributes to better performance of SRPNER.

The new version of this system recognizes more variations for naming persons: distinguished persons and first names alone. Moreover, system distinguishes names used for men from those used for women (Krstev et al., 2015). Presented results show that the system was more successful in rec-

ognizing names of men than women.

The output of the system are texts with XML tags for recognized entities inserted in them. Since the system allows embedded NEs, recognized names of persons can be components of other NEs, e.g. organizations.

2.2 The Preparation of the Gold Standard

The system presented in the previous section was used for the preparation of the gold standard – a large text sample annotated with personal names dubbed GOLDPERS. The sample consists of short news published on the Web in the period 2009–2016 by 4 Serbian daily newspapers (*Politika*, *Danas*, *Blic*, *Novosti*), one news portal (*B92*) and one weekly magazine (*Bazar*). The sample consists of 321,127 tokens (simple running words).

The forms of personal names taken into account and their tagging are presented in Table 1. The gold standard was produced following these steps:⁴

- Each text was annotated using SRPNER;
- Tags that did not refer to personal names were deleted;
- The remaining tags were evaluated as correct, partially correct (overlapping), not correct (not a name);
- The missing tags were inserted, and typos that led to incorrect tagging were corrected.

For some texts this process was repeated from one to four times which yielded “four levels” of gold standard. Between these repeated runs the development of SRPNER continued, as well as the enhancement of e-dictionaries of Serbian.

3 Training Different NER Systems

3.1 Training Sets

The gold standard GOLDPERS contains 9,046 sentences, each one enclosed in `<seg>...</seg>` tag. Named entities in sentences are annotated using tags listed in Table 1. These tags contain different levels of information: name type, role, gender. We wanted to examine the recognition of NEs on different level of details. Therefore, on the basis of the gold standard, we developed its four versions by

⁴The evaluation was performed as a homework by several generations of students of Library and Information Sciences at the Faculty of Philology, University of Belgrade, in the scope of the course Information Retrieval. Their work was checked by their professor, which means that texts were twice evaluated in each run.

type	example	original/translation
full	<m.persName.full>Mohamed El Baradei</m.persName.full> <f.persName.full>Ketrin Ešton</f.persName.full>	<i>Mohamed ElBaradei</i> <i>Catherine Ashton</i>
last	<m.persName.last>Obami</m.persName.last> <f.persName.last>Timošenkove</f.persName.last>	<i>Obama</i> <i>Tymoshenko</i>
first	<m.persName.first>Džim</m.persName.first> <f.persName.first>Tamara</f.persName.first>	<i>Jim</i> <i>Tamara</i>
role	<role>generalnog sekretara</role> <m.persName.full>Bana Ki Muna </m.persName.full> <role>Komesarka UN za ljudska prava</role> <f.persName.full>Nejvi Pilaj</f.persName.full>	<i>Secretary General</i> <i>Ban Ki-moon</i> <i>UN Commissioner for</i> <i>Human Rights Navi Pillay</i>
spec	<role>papa</role> <m.persName.first>Franja</m.persName.first> <role>kraljice</role> <f.persName.first>Viktorije</f.persName.first>	<i>Pope</i> <i>Francis</i> <i>Queen</i> <i>Victoria</i>

Table 1: Examples of types of personal names and their tags; M - masculine, F - feminine personal names

defining different mappings of XML tags. These mappings are named and represented in Table 2. Different versions of GOLDPERS are illustrated with examples presented in Table 3.

We split each of these four versions of the gold standard (namely PERS_{1, 3, 4, 9}) into training and test sets (containing 8, 151 and 895 sentences, respectively). We named this gold test set STUDENTS-GOLD.

As we wanted to have an independent text of the similar structure and content we prepared an additional set of news articles from *Danas* daily journal. This was one of the source journals for the STUDENTS-GOLD, but for this new sample, we have randomly chosen recent news (from year 2018, that is 2 years after the most recent news in GOLDPERS). This set of articles containing 860 sentences was tagged with SRPNER and manually corrected, thus producing the second test set DANAS-GOLD. The distribution of NE tags in the training set, and both test sets is given in Table 4.

We used four versions of the gold standard to train two different Named Entity Recognition systems: SPACY NER (Subsection 3.2) and STANFORD NER (Subsection 3.3). Trained models for Serbian are available on NER&BEYOND platform, which is presented in Section 6.

3.2 spaCy NER

spaCy (Honnibal and Montani, 2017) is a free, open-source library for advanced Natural Language Processing in Python. For different natural languages, it is able to perform tokeniza-

tion, POS-tagging, dependency parsing, lemmatization, sentence boundary detection, named entity recognition, similarity comparing, text classification, etc. It offers statistical models for a variety of languages, which can be installed as individual Python modules. It supports training new language models, as well.

We used it for training NER on our four versions of GOLDPERS. We coded a Python script that transforms each sentence into a training sample, represented as a list of triplets.⁶ For example, for the sentence “srpski reditelj Aleksandar Saša Petrović” (*Serbian director Aleksandar Saša Petrović*), the corresponding triplet representation for the PERS_4 model would be:

(0, 14, "ROLE"), (16, 39, "PERS_FULL")

where the first and the second element represent the start and the end character offset, while the third element represents the NE itself. Each of the four models were trained in 10 iterations, using 0.5 value for the drop-out parameter. These trained models can be inspected online.⁷

3.3 Stanford NER

STANFORD NER (Manning et al., 2014) is a Java implementation of a Named Entity Recognizer by the Stanford Natural Language Processing group. It is also known as *CRFClassifier*, since

⁶Training NER in spaCy,
<https://spacy.io/usage/training#ner>

⁷Visualization of SPACY NER for Serbian,
<http://ner.jerteh.rs/>

	PERS_1	PERS_3	PERS_4	PERS_9
m.persName.full	PERS	PERS_FULL	PERS_FULL	PERS_FULL_M
f.persName.full				PERS_FULL_F
x.persName.full				PERS_FULL_X
m.persName.first		PERS_FIRST	PERS_FIRST	PERS_FIRST_M
f.persName.first				PERS_FIRST_F
x.persName.first				PERS_FIRST_X
m.persName.last		PERS_LAST	PERS_LAST	PERS_LAST_M
f.persName.last				PERS_LAST_F
x.persName.last				PERS_LAST_X
role	-	-	ROLE	-

Table 2: Mappings of NE tags GOLDPERS to set tags used for training⁵

PERS_1	Film "Mančester na moru" <PERS>Keneta Lonergana</PERS> je u konkurenciji za šest "Oskara", dok je <PERS>Izabel Iper</PERS>, glavna junakinja filma "Ona" Holanđanina <PERS>Pola Ferhufena</PERS> nominovana za najbolju žensku ulogu.
PERS_3	Film "Mančester na moru" <PERS_FULL>Keneta Lonergana</PERS_FULL> je u konkurenciji za šest "Oskara", dok je <PERS_FULL>Izabel Iper</PERS_FULL>, glavna junakinja filma "Ona" Holanđanina <PERS_FULL>Pola Ferhufena</PERS_FULL> nominovana za najbolju žensku ulogu.
PERS_4	Film "Mančester na moru" <PERS_FULL>Keneta Lonergana</PERS_FULL> je u konkurenciji za šest "Oskara", dok je <PERS_FULL>Izabel Iper</PERS_FULL>, <ROLE>glavna junakinja filma "Ona"</ROLE> <ROLE>Holanđanina</ROLE> <PERS_FULL>Pola Ferhufena</PERS_FULL> nominovana za najbolju žensku ulogu.
PERS_9	Film "Mančester na moru" <PERS_FULL_M>Keneta Lonergana</PERS_FULL_M> je u konkurenciji za šest "Oskara", dok je <PERS_FULL_F>Izabel Iper</PERS_FULL_F>, glavna junakinja filma "Ona" Holanđanina <PERS_FULL_M>Pola Ferhufena</PERS_FULL_M> nominovana za najbolju žensku ulogu.

Table 3: The same sentence in four versions of the gold standard – PERS_{1, 3, 4, 9}

it is based on Conditional Random Fields (Lafferty et al., 2001). For training this model,⁸ we had to transform our texts into CoNLL02 IOB format (namely, “inside - outside - beginning”) with *conll* extension (Sang, 2002). For this purpose, we used XML \mapsto CoNLL converter available within NER&BEYOND on-line tool. An example of this format is given in Table 5.

4 Evaluation

In order to evaluate three NER system for personal names in Serbian, we need to have the output results in the same format. After running SPACY NER on a text, an output is provided in BRAT format with *ann* extension. This format is similar to the one for spaCy training, described in subsection 3.2. For the example given in the same sub-

⁸Training Stanford NER, <https://nlp.stanford.edu/software/crf-faq.shtml#a>

section, an output file has the following content:

```
T1 ROLE 0 14 srpski reditelj
T2 PERS_FULL 16 39 Aleksandar Saša Petrović
```

After running STANFORDNER on a text, an output is provided in already mentioned CoNLL02 format. We used CoNLL02 \mapsto BRAT converter available within NER&BEYOND online tool.

Finally, for both SPACY NER and STANAFORDNER output files, we applied ANN + TEXT \mapsto XML converter offered by Gemini, also available within NER&BEYOND online tool. An output of SRPNER is already an XML file with marked named entities, as is the gold standard explained in Section 2 and illustrated in Table 3.

We evaluated three NER systems using the open source Gemini tool, described in Section 6, that offers various options for files comparison (Feng, 2018). It is possible to select matching type: *strict*,

	Entity	Train	S	D
P_1	PERS	6312/2825	901/471	936/414
P_3	FULL	3280/1865	461/311	433/266
	FIRST	360/185	54/36	21/16
	LAST	2672/785	386/124	482/136
P_4	FULL	3280/1865	461/311	433/266
	FIRST	360/185	54/36	21/16
	LAST	2672/785	386/124	482/136
	ROLE	2069/1410	266/198	269/220
P_9	FULL _m	2732/1506	389/253	356/223
	FULL _f	547/358	72/59	75/42
	FULL _x	1/1	0	2/2
	FIRST _m	253/117	34/21	15/14
	FIRST _f	107/69	20/15	6/6
	FIRST _x	0	0	0
	LAST _m	2460/686	358/106	409/115
	LAST _f	192/93	27/17	73/21
	LAST _x	20/20	1/1	0

Table 4: Number of NE tags vs. number of different name forms in the training set and in test sets STUDENTS-GOLD (S) and DANAS-GOLD (D)

where exact overlapping of NE annotations is subsumed (both annotation labels are the same) or *weighted*, where partial overlapping is taken into account, but with some weighted value to measure overlapping segment. To indicate alignment type, one can choose among the two options: the first option is *greedyMatching*, where the matching of annotations in the first and second files is done with a greedy algorithm that tries to match the closest annotations first. The second option is *maxMatching*, where the matching of annotations in the first and second file is done optimally using a maximum matching algorithm in bipartite graphs. An annotation in the first file will correspond to at most one annotation in the second file, and vice versa, in both cases.

We run $2 \times 3 \times 4$ evaluation rounds: two test sets, three NERs and four models per each. All trials were run with *strict* matching type and *maxMatching* alignment type.

To indicate the chosen score type to evaluate the correspondence between one annotation from the first file and one annotation from the second file, calculation of precision P , recall R and F -measure in Gemini comes in three different flavors:

weak an annotation of the first file will be considered as corresponding to an annotation of

Fascinirala	O
me	O
je	O
Sonja	B-PERS
Savić	I-PERS
svojim	O
transformacijama,	O
Anica	B-PERS
Dobra	I-PERS
šarmom...	O

Table 5: CoNLL02 IOB format – the beginning of the sentence *I was fascinated by Sonja Savić and her transformation, Anica Dobra and her charm...*

the second file if they intersect on at least one character;

strict an annotation of the first file will be considered as corresponding to an annotation of the second file if they start and end exactly at the same characters;

weighted the match is scored by the ratio of the number of characters common to both annotations divided by the total number of characters covered by at least one of the two annotations.

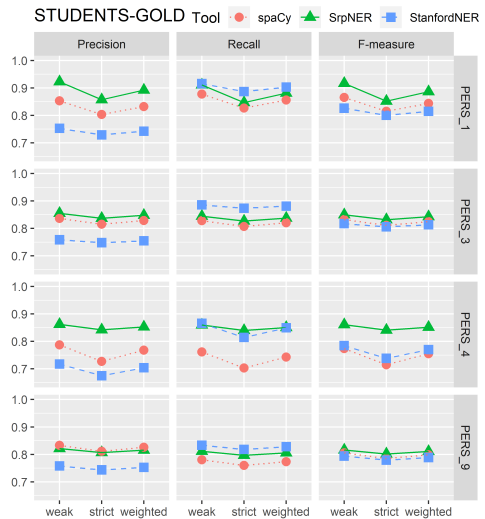


Figure 1: The evaluation of SPACY NER, SRPNER and STANFORD NER on STUDENTS-GOLD

5 Discussion

The results of three NER systems, four models and two test texts are presented in Table 6. The results show that in all cases (except one) SRPNER achieved the best precision, in all cases (except one) STANFORD NER achieved the best recall,

		SPACY NER			SRPNER			STANFORD NER		
model		<i>P</i>	<i>R</i>	<i>F</i> ₁	<i>P</i>	<i>R</i>	<i>F</i> ₁	<i>P</i>	<i>R</i>	<i>F</i> ₁
STUDENTS	PERS_1	0.804	0.827	0.815	0.857	0.847	0.852	0.729	0.887	0.800
	PERS_3	0.815	0.807	0.811	0.837	0.827	0.832	0.748	0.873	0.806
	PERS_4	0.727	0.703	0.715	0.842	0.840	0.841	0.675	0.815	0.738
	PERS_9	0.812	0.760	0.785	0.807	0.797	0.802	0.744	0.818	0.779
DANAS	PERS_1	0.819	0.870	0.844	0.916	0.841	0.877	0.790	0.902	0.842
	PERS_3	0.864	0.854	0.859	0.905	0.830	0.866	0.791	0.874	0.830
	PERS_4	0.807	0.792	0.799	0.907	0.824	0.863	0.716	0.825	0.767
	PERS_9	0.818	0.794	0.806	0.872	0.800	0.835	0.761	0.808	0.784

Table 6: The comparison of strict precision, recall and F_1 between NER systems, models and test sets.

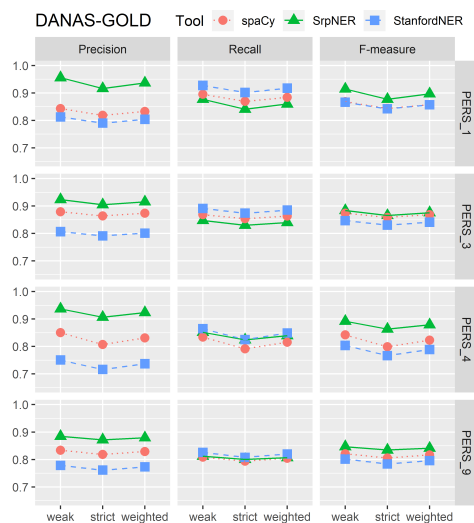


Figure 2: The evaluation of SPACY NER, SRPNER, and STANFORD NER on DANAS-GOLD

while SRPNER achieved the highest F_1 measure in all cases. STANFORD NER and SRPNER performed better on both test texts with models that use less tags (PERS_1 and PERS_2), while SRPNER performed significantly worse only for the model with 9 tags. Contrary to our expectations, all NER systems for all models achieved better results for the independent test text DANAS_GOLD than for the test set randomly chosen from the gold standard. Namely, a number of news in GOLD-PERS that come from 6 different sources were produced at the same time period, and thus involved same persons. However, that did not influence results favorably towards STUDENTS_GOLD test text.

All measures are for all NER systems and models highest for weak calculation, followed by weighted, the strict calculation giving the lowest result. However, as displayed in Figure 1 for the

STUDENTS-GOLD test set and in Figure 2 for the DANAS-GOLD test set the mutual relationship between three NER system remains the same.

We also compared performance of all three NER systems by each named entity type (Figure 3). Results for all three models distinguishing entity types show that all three systems perform poorly in recognizing first names only. For STANFORD NER and SPACY NER it can be explained by the considerably smaller number of these tags in training texts compared to other tags (see Table 4). As for SRPNER one can presume that developers devoted less effort to this entity type occurring only occasionally in newspaper texts. Similarly, in all experiment settings, the recognition of full names was better than the recognition of last names only. Again, the number of last name tags was smaller than the number of full name tags (Table 4). A rule based system SRPNER makes use of a personal name context in cases of disambiguity which tends to be less specific in the case of the use of a last name only.

One can also note that when using the model PERS_4, SRPNER system performs best in recognizing the role entity. Between other two systems, STANFORD NER achieves better recall and SPACY NER slightly better precision.

The chart for model PERS_9 shows that all systems according to F_1 measure recognize better masculine names than feminine names regardless of entity types. Feminine names show greater variety of forms than masculine names, especially when only last names are used; moreover, they occur significantly less than masculine names in newspaper texts (as pointed in (Krstev et al., 2015) there is approximately one feminine name per 7 seven masculine names) and confirmed in our training sets (Table 4).

We compared the performance of SRPNER with its previously reported results. In (Krstev et al., 2005) results for the recognition of personal names were $P = 0.97$, $R = 0.86$, $F_1 = 0.91$. One notes that all measures are higher than those obtained when using the GOLDPERS (see rows PERS_1 in Table 4), which can partly be due to the inclusion of first names only into the present version of SRPNER. On the other hand, results obtained for the recognition of roles with GOLDPERS ($F_1 = 0.87$ for STUDENT and $F_1 = 0.86$ for DANAS) were higher than those presented in the same paper ($F_1 = 0.83$). The same goes for the recognition of last names only: the previous result was $F_1 = 0.78$, while the use of SRPNER on the gold standard yielded $F_1 = 0.86$ for STUDENT and $F_1 = 0.85$ for DANAS.

The capability to distinguish masculine and feminine names was compared to the results presented in (Krstev et al., 2015). The results obtained in presented experiments were lower for all NE types: masculine full $F_1 = 0.97$ vs. $F_1 = 0.90/0.94$ (STUDENT/DANAS), feminine full $F_1 = 0.94$ vs. $F_1 = 0.86/0.86$, masculine last $F_1 = 0.89$ vs. $F_1 = 0.85/0.83$, feminine last $F_1 = 0.79$ vs. $F_1 = 0.49/0.64$. One can presume that the use of gold standard produces more reliable results.

To the best of our knowledge, STANFORD NER and SPACY NER were used for the first time for the recognition of personal names in Serbian texts. Ljubešić et al. (2013) used STANFORD NER to build models for Croatian and Slovene. When they used distributional similarity to improve results, on texts coming from different sources they obtained the following results: for Croatian $P = 0.91$, $R = 0.93$ and $F_1 = 0.92$, higher than STANFORD NER for the model PERS_1, and for Slovene $P = 0.82$, $R = 0.87$ and $F_1 = 0.84$, comparable with STANFORD NER for the model PERS_1 (Table 6). One should note, however, that their test set contained a smaller number of personal names (approximately one third of number of personal names in our test sets).

Jiang et al. (2016) compared 4 NER systems, two of which were STANFORD NER and SPACY NER, for English. Their test set consisting of Wiki articles contained approximately the same number of personal names as our both sets – around 900. Their results were for STANFORD NER $P = 0.72$, $R = 0.87$ and $F_1 = 0.79$, and for SPACY NER

$P = 0.73$, $R = 0.73$ and $F_1 = 0.73$. Our results are comparable in the sense that they also show that STANFORD NER achieves the best recall, while SPACY NER tends to have the more balanced precision and recall.

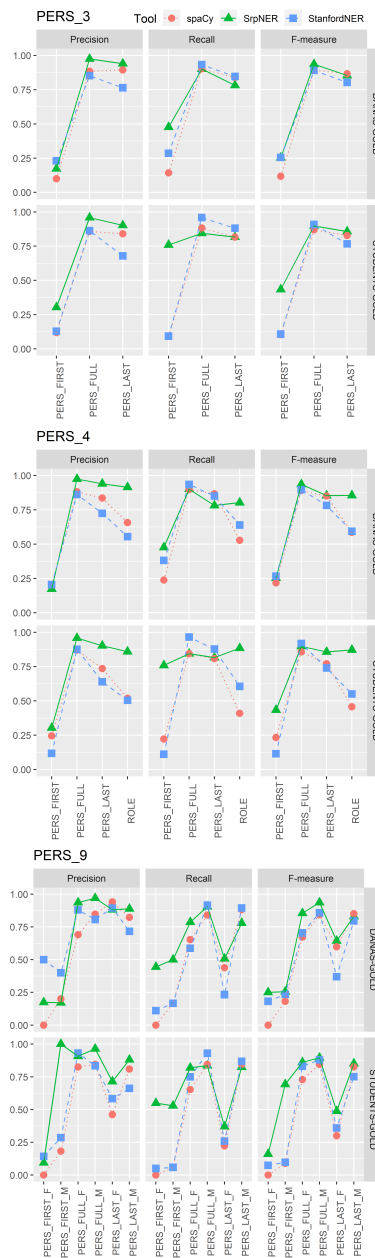


Figure 3: Evaluation of SRPNER, SPACY NER and STANFORD NER on two test sets, by each named entity type

6 Online Tool for NER

Serbian NER team (2019) offers an on-line tool for different purposes related to Named Entity Recognition. First, it supports conversion among different formats common for representations of NERs:

BRAT, CoNLL02 and XML. BRAT (Stenetorp et al., 2012) is a web-based tool⁹ for text annotation, i.e., for adding notes to existing text documents. It is designed for structured annotation, allowing embedded annotations, which are especially convenient for NER. Annotations are external, so for each text file, an additional annotation file contains annotation data described in Section 4. CoNLL02 is a two-column format, also described in Section 4. An example of XML file whit tags interpreted as NEs is given in Table 3.

NER&BEYOND contains nine different modules:

XML \mapsto **BRAT** module supports transformation of XML files which tags are interpreted as named entities, to BRAT format;

BRAT \mapsto **XML** module supports transformation of files in BRAT format and their corresponding textual files to XML format;

BRAT \mapsto **CoNLL02** module supports transformation of files in BRAT format and their corresponding textual files to CoNLL02 format, using a Python script that is a part of BRAT package;

CoNLL02 \mapsto **BRAT** module supports transformation of files in CoNLL02 format to BRAT format and their corresponding textual files, using a Python script that is a part of BRAT package;

XML \mapsto **CoNLL02** module supports transformation of XML files which tags are interpreted as named entities, to CoNLL02 format;

spaCy NER module provides NE annotation using spaCy (Honnibal and Montani, 2017), a free, open-source library for advanced NLP tasks in Python. This portal offers automatic annotation of texts in English, Spanish, German, Portuguese, French, Italian, Dutch and Serbian;

StanfordNER module provides Named Entity annotation using STANFORD NER models (Manning et al., 2014), which are available for Serbian, English and German with different levels of details, e.g. number of NE classes. Serbian model is developed withing presented research, while English and German are integrated from Stanford repository;

NER statistics module is developed for analysis of annotated text collections in BRAT, that can be automatically downloaded via BRAT

web interface. Various statistics related to distributions of named entities and attributes can be computed, including frequencies of annotated entities, classes, attributes per document and collection;

Gemini tool allows comparison of two text annotation files and provides different alignment scores. It is possible to compare a pair of XML files, a pair of files in BRAT format and one XML file against a file in BRAT format. The first file is the output of a NER system and the second file represents a gold standard.¹⁰

7 Future Work

For the upcoming research, we plan to apply the procedure we used for personal names to other NE classes (organization, location, event, temporal, quantitative, etc) and to experiment with other ML NER methods and tools with an ultimate goal to produce a successful hybrid system. The important next step is the enhancement of our newspaper corpus with other types of text (Wikipedia articles, domain texts, literary texts). The literary texts would be particularly important for improving the recognition of first names. Finally, another intended step is Entity Linking (EL), i.e. disambiguation of recognized named entities to a knowledge base, such as Wikidata, DBpedia WordNet and BabelNet. Such example would be automatically assigning Wikidata URL that points to a biography of a famous person to the corresponding named entity detected in text.

Acknowledgments

This research was supported by Serbian Ministry of Education and Science under the grants #III 47003 and 178006.

References

- Yuheng Feng. 2018. Gemini - un module de comparaison de deux fichiers de textes annotés. Université Paris-Est Marne-la-Vallée, l'Institut Gaspard-Monge, <https://github.com/fyh828/gemini>.
- Jenny Rose Finkel and Christopher D. Manning. 2009. Nested Named Entity Recognition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume*

⁹BRAT, <https://brat.nlplab.org>

¹⁰Gemini, <https://github.com/fyh828/gemini/>

- I. Association for Computational Linguistics, pages 141–150.
- Nathalie Friburger and Denis Maurel. 2004. Finite-state Transducer Cascades to Extract Named Entities in Texts. *Theoretical Computer Science* 313(1):93–104.
- Ralph Grishman and Beth Sundheim. 1996. Message Understanding Conference-6: A Brief History. In *Proceedings of the 16th International Conference on Computational Linguistics (COLING 1996)*, volume 1.
- Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing. To appear, <https://pypi.org/project/spacy/2.0.16/>.
- Martin Jansche and Steven Abney. 2002. Information Extraction from Voicemail Transcripts. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*.
- Ridong Jiang, Rafael E Banchs, and Haizhou Li. 2016. Evaluating and Combining Name Entity Recognition Systems. In *Proceedings of the Sixth Named Entity Workshop, joint with 54th ACL*, pages 21–27.
- 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. <https://doi.org/10.1093/logcom/exs079>.
- Cvetana Krstev, Miloš Utvić, and Jelena Jaćimović. 2015. Ako koza laže, rog ne laže - gde su i ko su žene u srpskoj dnevnoj štampi [Facts Are Stubborn Things – Women in Serbian Daily Press]. *Književstvo - časopis za studije književnosti, roda i kulture V*. <https://doi.org/10.18485/knjiz.2015.1.24>.
- Cvetana Krstev, Duško Vitas, and Sandra Gucul. 2005. Recognition of Personal Names in Serbian Texts. In *International Conference Recent Advances in Natural Language Processing (RANLP'05)*, pages 288–292.
- George R. Krupka and Kevin Hausman. 1998. Isoquest: Description of the netowlTM extractor system as used in MUC-7. In *Message Understanding Conference (MUC-7)*.
- John Lafferty, Andrew McCallum, and Fernando C.N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of ICML*, pages 282–289.
- Nikola Ljubešić, Marija Stupar, Tereza Jurić, and Željko Agić. 2013. Combining Available Datasets for Building Named Entity Recognition Models of Croatian and Slovene. *Slovenščina* 2(1):35–57.
- Christopher Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.
- Denis Maurel, Nathalie Friburger, Jean-Yves Antoine, Iris Eshkol, and Damien Nouvel. 2011. Cascades de transducteurs autour de la reconnaissance des entités nommées. *Traitement automatique des langues* 52(1):69–96.
- David Nadeau and Satoshi Sekine. 2007. A Survey of Named Entity Recognition and Classification. *Linguisticae Investigationes* 30(1):3–26.
- E.F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 Shared Task: Language-independent Named Entity Recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.
- Satoshi Sekine, Kiyoshi Sudo, and Chikashi Nobata. 2002. Extended Named Entity Hierarchy. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*. European Language Resources Association (ELRA), Las Palmas, Canary Islands - Spain. <http://www.lrec-conf.org/proceedings/lrec2002/pdf/120.pdf>.
- Serbian NER team. 2019. NER&Beyond. <http://nerbeyond.jerteh.rs/>.
- Sameer Singh, Dustin Hillard, and Chris Leggetter. 2010. Minimally-supervised Extraction of Entities from Text Advertisements. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, pages 73–81.
- Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Junichi Tsujii. 2012. BRAT: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations Session at EACL 2012*.
- Duško Vitas and Cvetana Krstev. 2012. Processing of Corpora of Serbian using Electronic Dictionaries. *Prace Filologiczne* LXIII:279–292.