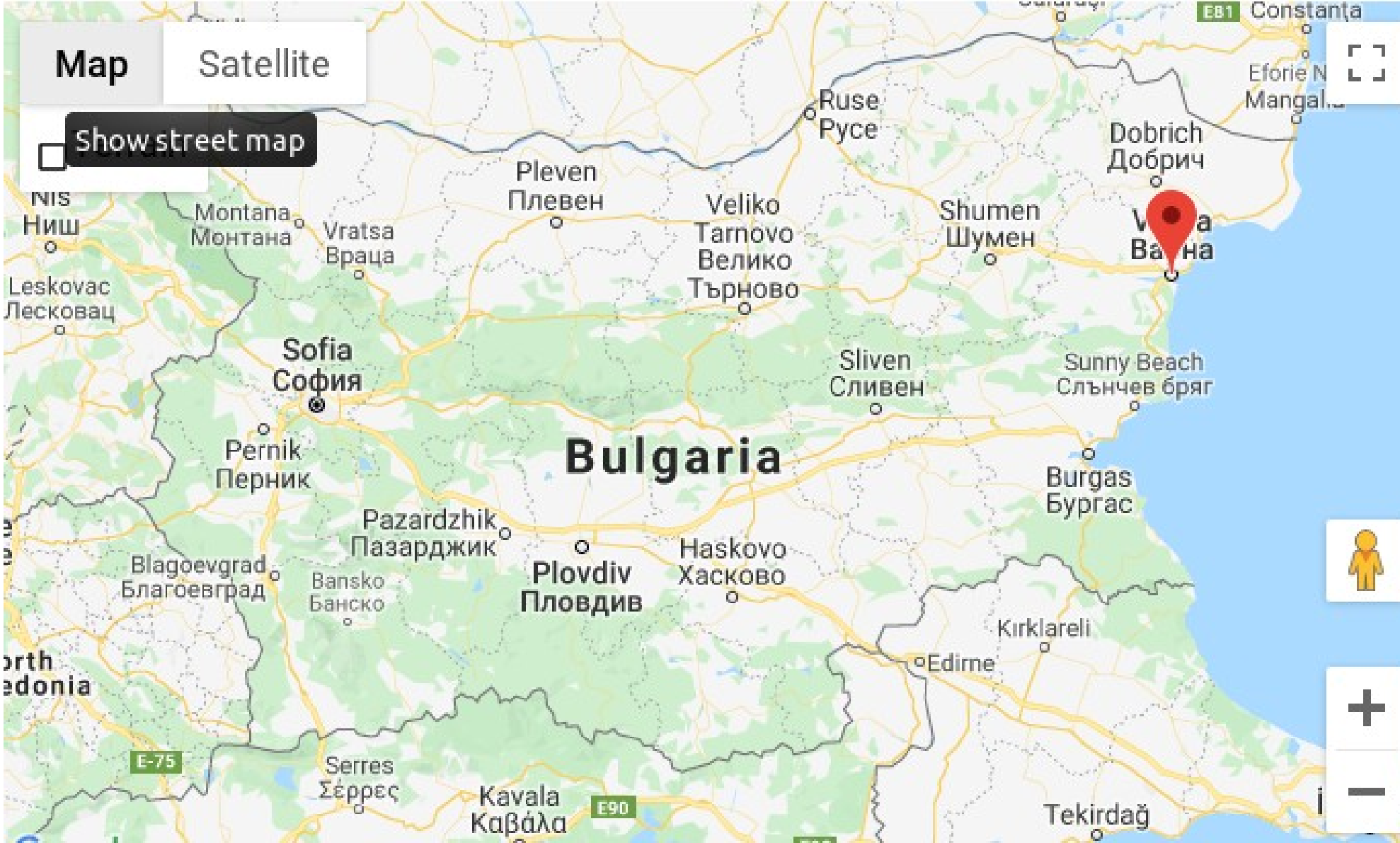
Recent Advances in Natural Language Processing

2 – 4 September, 2019, Varna, Bugarska

Branislava Šandrih

24. X 2019

JeRTeh seminar, Beograd, Srbija


Map
Satellite
Show street map
Bulgaria
Sofia
София
Pernik
Перник
Pazardzhik
Пазарджик
Plovdiv
Пловдив
Blagoevgrad
Благоевград
Bansko
Банско
Haskovo
Хасково
Pleven
Плевен
Ruse
Русе
Veliko
Tarnovo
Велико
Търново
Shumen
Шумен
Dobrich
Добрич
Sliven
Сливен
Sunny Beach
Слънчев бряг
Burgas
Бургас
Kırklareli
Edirne
Tekirdağ
Serres
Σέρρες
Kavala
Καβάλα
Montana
Монтана
Vratsa
Враца
Leskovac
Лесковац
E-75
E90
E81


GREE

Organizatori

- Research Group in CL, University of Wolverhampton, UK
  - PC CHAIR: Prof Dr Ruslan Mitkov
- LMD, Institute of Information and Communication Technologies, BAS, BG
  - OC CHAIR: Prof Dr Galia Angelova

	CHERNO MORE hall	**VARNA** hall	**ODESSOS** hall	**Small hall** 2nd floor
29 August		**DLinNLP Summer School**		
30 August		**DLinNLP Summer School**		
31 August		**RANLP Tutorials**		
1 Sept.		**RANLP Tutorials**		
2 Sept.	**RANLP Conference**	**RANLP Conference**	**RANLP Conference**	
3 Sept.	**RANLP Conference**	**RANLP Conference**	**RANLP Conference**	
4 Sept.	**RANLP Conference**	**RANLP Conference**	**RANLP Conference and Student Research Workshop**	
5 Sept. **Post-conference events**	**HiT-IT 2019** Second Workshop on Human-Informed Translation and Interpreting Technology **Welcome reception** post-conference events	**International Conference** Biographical Data in a Digital World 2019	**BUCC 2019** 12th Workshop on Building and Using Comparable Corpora	**Workshop** LT for digital historical archives with a special focus on CEE and SEE, Middle East and North Africa
6 Sept. **Post-conference events**	**HiT-IT 2019** Second Workshop on Human-Informed Translation and Interpreting Technology	**International Conference** Biographical Data in a Digital World 2019	**Multiling 2019** Workshop on Summarization Across Languages, Genres and Sources	

Društveni događaji

- Summer School on Deep Learning in Natural Language Processing welcome cocktail reception Thursday, 29 August in the evening.
- Main conference and tutorials welcome cocktail reception, Sunday, 1 September in the evening.
- RANLP'2019 Gala Dinner Monday, 2 September (Folklore and entertainment programme during the dinner; dances after the dinner)
- The Conference Excursion on Tuesday, September 3 in the afternoon. The Conference Excursion will be a bus trip to two major historical locations in Bulgaria: the Madara Rider and the Pliska fortress.
- Welcome cocktail reception for the post-conference events on Thursday, 5 September in the evening.

DlinNLP: Summer School on Deep Learning in Natural Language Processing

- **Tim Rocktäschel** (University College London)
  - Introduction to Deep Learning for NLP
- **Hinrich Schütze** (Ludwig Maximilian University, Munich)
  - Neural-based representation Learning
- **Kyunghyun Cho** (New York University)
  - Latest topics in representation learning for language
- **Marek Rei** (University of Cambridge)
  - Application of Deep Learning in NLP

DLinNLP 2019

RANLP-2019 SUMMER SCHOOL ON DEEP LEARNING IN NLP

VARNA hall

	29 August 2019	30 August 2019
Morning Coffee break: 10:30 - 11	09:00 - 10:30 **Tim Rocktäschel** *Introduction to Deep Learning* 11:00 - 12:30 **Hinrich Schütze** *Neural Representation Learning*	09:00 - 12:30 **Kyunghyun Cho** *Latest topics in Representation Learning*
Lunch	Lunch break 12:30-14:00	Lunch break 12:30-14:00
Afternoon Coffee break: 15:30 - 16	14:00 - 15:30 **Hinrich Schütze** *Neural Representation Learning* 16:00 - 19:00 **Heike Adel** Practical Session *General introduction to PyTorch Hands-on material on embeddings and how pre-trained embeddings can improve models for NLP* 19:00-20:00 Informal poster session	14:00 - 15:00 **Alexander Popov** Practical Session *Sense/synset embeddings and graph enrichment methods* 15:00 - 17:00 **Omid Rohanian** and **Shiva Taslimipoor** Practical Session *Sequence labelling and tagging* 17:30 - 19:00 **Marek Rei** *Application of Deep Learning in NLP*
	20:30 Main Hotel Lobby **Welcome Reception**	

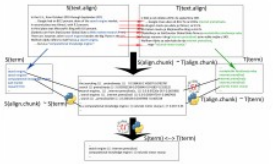
Impact of Text Classification in Natural Language Processing Applications

Branislava Šandrih, PhD student of Computer Science
University of Belgrade, Faculty of Mathematics, Belgrade, Serbia
Web applications and other tools: http://portal.jerteh.rs/

Extraction and Validation of Bilingual Terminology Pairs

(Krstev et al. 2018)

We describe an approach for obtaining bilingual terminology pairs automatically, and demonstrate its effectiveness for English/Serbian language pair:



The compiled list of bilingual MWTs is afterwards manually evaluated. A binary classifier that maps pairs (samples) to a space of linguistic features (Chart 1) was trained to classify candidate translation pairs in the future. Pairs were classified either as good or as bad translations. This classifier (Table 1) can be used for compilation of the domain lexica for the same domain in the future.

CLF	Accuracy	F1 score	Precision	Recall
NB	70.89 ± 0.68	75.96 ± 0.81	64.93 ± 0.68	79.68 ± 0.79
LR	78.44 ± 0.66	82.08 ± 0.67	77.16 ± 0.61	87.90 ± 0.55
SVM	63.85 ± 15.24	75.07 ± 5.73	71.19 ± 6.24	64.86 ± 25.17
RBF	78.49 ± 0.80	82.89 ± 0.73	77.39 ± 0.78	87.65 ± 0.84
RF	77.29 ± 0.68	79.96 ± 0.67	79.73 ± 0.80	78.25 ± 1.07
GB	78.25 ± 0.79	81.75 ± 0.80	76.86 ± 0.76	87.24 ± 0.84

Table 1. Classification results

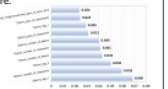


Chart 1. Feature importance

Authorship Identification of Short Texts

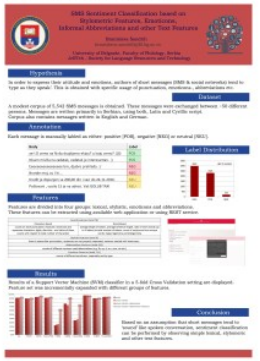
(Šandrih 2018)

Is it possible to tell who is the author of a short text (e.g. SMS) based on linguistic and stylometric features of the text?

Sentiment Classification in Short Messages

(Šandrih 2019)

Is it possible to tell which sentiment is contained in a short message (e.g. SMS) based on linguistic and stylometric features of the message?



Classifying Good Dictionary Examples for Serbian

(Stanković et al. 2019)

We present a model for selection of good dictionary examples (GDEX) for Serbian and development of initial model components. The proposed method is based on a thorough analysis of various lexical and syntactic features in a corpus compiled of examples from the five digitized volumes of the Serbian Academy of Sciences and Arts (SASA) dictionary.

For each (dictionary entry, example) pair, we extracted the following set of features: sentence length, number of digits, number of "weird" characters, punctuation count, number of all tokens, average token length, maximum token length, number of capitalised words, number of rare tokens (in contrast to the referent corpus), average frequency of words and number of pronouns.

The feature distribution of examples from this corpus is compared with feature distribution of sentence samples extracted from corpora comprising various texts. The analysis showed that there is a group of features which are strong indicators that a sentence should not be used as an example. The remaining features, including detection of non-standard and other marked lexis from SASA dictionary, are used for ranking. The selected candidate examples, represented as feature-vectors, are used for GDEX ranker for Serbian candidate examples and a supervised machine learning model (Table 2) for classification on standard and non-standard Serbian sentences, for further integration into a solution for present and future dictionary production projects.

	Precision	Recall	F1-score	Number of samples
NO	0.84	0.68	0.75	11,056
OK	0.73	0.87	0.79	11,180
Total	0.78	0.77	0.77	22,236

Table 2. Results of the Logistic Regression binary classifier

Contact

Branislava Šandrih
MSc of Mathematics
Junior lecturer at Faculty of Philology
University of Belgrade, Serbia
http://branislava.github.io
branislava.sandrih@fil.bg.ac.rs

References

[Stanković et al. 2019] Ranka Stanković, Branislava Šandrih, Cvetana Krstev, Duško Vitas, and Aleksandra Matković. SASA Dictionary as the Gold Standard for Good Dictionary Examples for Serbian. In eLex 2019 Smart Lexicography, 2019.

[Šandrih 2019] Branislava Šandrih. SMS Sentiment Classification based on Lexical Features, Emoticons and Informal Abbreviations. Yerdua Journal of Computing, 2019.

[Krstev et al. 2018] Cvetana Krstev, Ranka Stanković, and Mihailo Matković. Using English Bells Text at Serbian Multi-Word Terminology. In Proceeding of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), Paris, France, May 2018. European Language Resources Association (ELRA).

[Šandrih 2018] Branislava Šandrih. Fingerprints in SMS messages: Automatic Recognition of a Short Message Sender Using Gradient Boosting. In 5th

Tutorijali

- Preslav Nakov (Qatar Computing Research Institute, HBKU)
- Valia Kordoni (Humboldt University of Berlin)
- Antonio Miceli Barone (University of Edinburgh) & Sheila Castilho (Dublin City University)
- Vlad Niculae & Tsvetomila Mihaylova (Institute of of Telecommunications, Lisbon)

RANLP-2019 TUTORIALS, VARNA hall

	31 August 2019	**1 September 2019**
9:30-13:10	**Preslav Nakov** *Fact Checking: Truth Seeking in the Age of Disinformation* Coffee break: 10:30 - 10:50 Break: 11:50 – 12:10	**Antonio Miceli Barone** and **Sheila Castilho** *Neural Machine Translation* Coffee break: 10:30 - 10:50 Break: 11:50 – 12:10
Lunch	Lunch break 13:10-14:30	Lunch break 13:10-14:30
14:30-18:10	**Valia Kordoni** *Deep Learning for Metaphors and Idioms* Coffee break: 15:30 - 15:50 Break: 16:50 – 17:10	**Vlad Niculae** and **Tsvetomila Mihaylova** *Latent Structure Models for NLP* Coffee break: 15:30 - 15:50 Break: 16:50 – 17:10
		19:30 VARNA hall **RANLP Welcome Cocktail** and **Tutorial in Bulgarian Folk Dances**

Glavna konferencija

- Oko 220 poslatih radova
- 150 prihvaćenih radova
  - 18 dugih radova
  - 37 kratkih radova
  - 95 postera
- Bez razlike u zborniku

Izlagači po pozivu

- Hinrich Schütze
  - Teaching Deep Networks Lexical Semantics: the Easy Way or the Hard Way?
- Kyunghyun Cho
  - A Generalized Framework of Sequence Generation

Izlagači po pozivu

- Ken Church
  - Setting Appropriate Expectations: are Deep Nets too Hot? Too Cold? Or just Right?
- Sebastian Padó
  - Entities as a Window into (Distributional) Semantics

Izlagači po pozivu

- Preslav Nakov
  - Detecting the ‘Fake News’ Before they were even Written

Naš rad (poslednji dan konferencije)

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

Branislava Šandrih, Cvetana Krstev & Ranka Stanković

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 Minning and Geology, Belgrade, Serbia
ranka@rgf.bg.ac.rs

Motivation

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 lexiconbased 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 F1), while Stanford models have the highest recall.

The produced models are incorporated into a Web platform NER&Beyond thatprovides various NE-related functions.

Gold Standard for Serbian NER

The first NER system for Serbian (SRPNER) was a rule- and lexicon-based system. It 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. Each transducer rely in its work on the results of previous transducers and on e-dictionaries of Serbian.

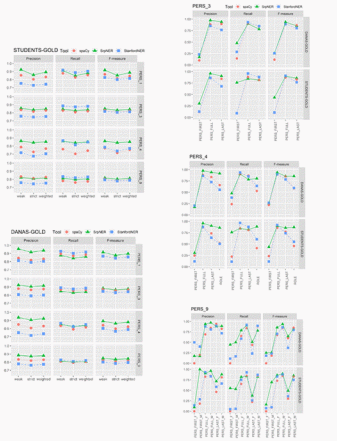
SRPNER recognizes 11 classes of NEs: dates (moments and periods), time (moments and periods), money expressions, measurement expressions, geopolitical names (countries, settlements, oronyms and hydronyms), and personal names (one or more last names with or without first names and nicknames).

SRPNER 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 by 4 Serbian daily newspapers, one news portal (B92) and one weekly magazine (Bazar). The sample consists of 321,127 tokens.

The gold standard was produced following these steps:

- *Each text was annotated using SRPNER;*
- *Tagsthatdidnotrefertopersonalnameswere 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.*

Results: 4 models × 3 NERs × 2 Test Sets



NER & Beyond, http://nerbeyond.jerteh.rs

An on-line tool for different purposes related to Named Entity Recognition was developed by Serbian NER team from *JeRTeh - Society for Language Resources and Technologies*. The tool consists of 9 modules.

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

BRAT→XML module supports transformation offiles in BRAT format and their corresponding textual files to XML format;

BRAT→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→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→CoNLL02 module supports transformation of XML files which tags are interpreted as named entities, to CoNLL02 format

NER statistics module is developed for analysis of annotated text co llections 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,a ttributes 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.

spaCyNER module provides NE annotation using spaCy, 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; 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. For example, for the sentence "srpski reditelj Aleksandar Sasa Petrovic" (Serbian director Aleksandar Sasa Petrovic), 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.

StanfordNER module provides Named Entity annotation using STANFORD NER models, 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;

Благодаря!