Named Entity Recognition

Thursday, June 6, 2019, 11:00 – 12:00

Cvetana Krstev, full-time professor
University of Belgrade

In this presentation we first introduce the notion of "named entity" that is being in the focus of researchers dealing with text processing, specifically information extraction, for several decades. We will than present various strategies and annotation schemes used for the annotation of named entities. Subsequently, we will present various methods used for the recognition of named entities in unstructured texts: heuristics-based methods, statistical and machine learning methods, and lexicon- and rule-based methods.

We will present in more details a rule-based named entity recognition system for Serbian, resources it is based on and strategy it uses. This strategy will be illustrated on the example of personal names that occur in texts in various forms. Finally, we will describe how this system was used to prepare a "gold standard", a collection of newspaper articles annotated with personal names. The purpose of this gold standard is to be used as a training corpus for machine-learning named entity recognition system.

NER&Beyond: Training and Evaluating Machine Learning Models

Thursday, June 6, 2019, 12:00 – 13:00

Branislava Šandrih,
Junior lecturer
University of Belgrade, Faculty of Philology

In this presentation, we explain how the gold standard in Serbian, annotated with personal names, was used for training two different ML-based NER models: spaCy and StanfordNER. We explain how we trained these models and which format conversions had to be performed.

After the training procedure, these two models were evaluated on two validation sets: one taken out from the gold standard, and an independent one, but with similar origin and structure. A comparison of these models with the existing rule- and lexicon-based NER for Serbian, SrpNER, was done afterwards. Since these three models give outputs in different formats (XML, CoNLL02 and BRAT), we recognized a need for a tool that supports conversion of different file formats common for representing named entities.

Threfore, we present an on-line platform, *NER&Beyond,* developed for various NER-related tasks: conversion between different file formats; named entity recognition, annotation and visualization using trained ML-models; tools for statistics and evaluation. The talk is concluded with a practical demo of the platform.

Bilingual Terminology Extraction

Thursday, June 6, 2019, 17:00 – 18:00
Cvetana Krstev, full-time professor
University of Belgrade

In this presentation we will present the importance of bilingual terminology, both for humans and for various machine applications. In the era of very dynamic technology development and the explosion of published information, it is very difficult to produce bilingual terminology resources exclusively manually. We will present some methods used for monolingual terminology extraction from domain texts: statistical, rule-based, lexicon-based and hybrid. For bilingual terminology extraction additional resources are needed, primarily a bilingual aligned domain corpus.

In this presentation we will present resources and tool that we used for our experiment in bilingual terminology extraction in the domain of library and information sciences: a bilingual domain corpus consisting of journal articles, a bilingual domain lexicon, a bilingual list of general lexica containing not only basic forms but also inflected forms, some open-source term extractors for English, and finally a lexicon- and rule-based term extractor for Serbian. All these resources are used by a bilingual terminology extraction system, dubbed BiLTE.

BiLTE: Automatic Alignment and Validation of Bilingual Pairs of MWEs

Thursday, June 6, 2019, 18:00 – 19:00
Branislava Šandrih, Junior lecturer
University of Belgrade, Faculty of Philology

In this talk we demonstrate a language-independent approach for automatic extraction, alignment and validation of bilingual terminology pairs, obtained from parallel corpora.

The first step in the procedure is a compilation of bilingual candidate translation pairs. For the purpose of word-alignment, phrase extraction and phrase scoring, we used GIZA++, a statistical machine translation toolkit that relies on a Hidden Markov Model. As an input, GIZA++ requires two sentence-aligned files in different languages, that represent translations of each other. As a result, a list of candidate pairs of text chunks is compiled.

After different filtering steps, a list of bilingual terminology pairs was compiled. These pairs were afterwards manually evaluated, as either good or bad translations. Having this list, we represented each sample, i.e. a pair of bilingual chunks, as a 178-dimensional feature vector. In the talk we describe the feature list and we offer a publicly-available web service for the extraction of these feature on a text.

This dataset of feature-vectors was used for evaluation of various binary Machine Learning classifiers in a 5-CV fold setting: Naive Bayes; Logistic Regression; Linear and Radial-Basis Support Vector Machines; Random Forests and Gradient Boosting.

Finally, we reveal an on-line platform for automatic compilation of bilingual terminology lists, *BiLTE.* The talk is concluded with a practical demo of the platform on an example of Serbian-English terminology from the domain of Library and Information Science.