UDC 004.056.5

M.I. Kozlenko, I.M. Lazarovych, M.V. Kuz
Vasyl Stefanyk Precarpathian National University
mykola.kozlenko@pnu.edu.ua
ihor.lazarovych@pnu.edu.ua
mykola.kuz@pnu.edu.ua

NAMED ENTITY RECOGNITION SOFTWARE FOR AUTOMATED EXTRACTION OF CALL SIGNS AND SERVICE MESSAGES IN SHORTWAVE RADIO COMMUNICATIONS

One of the key aspects for ensuring the national security is the timely detection of threats and intentions of other countries, as well as understanding the sources of the threats. For a long time, messages circulating in shortwave radio networks of military or governmental communications are an important information source and have special significance for such intent detection. The extraction of structured information from unstructured text messages is an important task that leads to creation of possibility for fully automated data processing. And it is not an easy task. Further, the structured data can be used for intent detection in machine learning models. The named entity recognition (NER) is a subtask of information extraction that seeks to locate and classify named entity mentioned in unstructured text into predefined categories [1]. The main challenges for NER for call signs and service/test words are their artificiality, unknown distributions, and extremely low size of available examples that can be used as training dataset.

The main goal of this research is to develop an NER model for extraction such entities as call signs and test words from text messages. Test words are used for estimating the readability [2] of radio transmission. As a rule, test words are out-of-vocabulary artificially synthesized entities.

The most popular NER platforms include GATE, OpenNLP, spaCy. As a rule, NER systems are created using linguistic grammar-based techniques as well as statistical models such as machine learning [3]. Many different classifier types are used to perform

machine-learned NER, with conditional random fields being a typical choice [4]. As main framework we have chosen the spaCy that is an open-source software library for advanced natural language processing, written in Python [5].

The dataset is based on texts collected from messages transmitted by one well-known shortwave radio station. The dataset contains 1380 test words. The entire data set is split into train (1000 records) and test (about 380 records) parts. About 10 % of train set is used as development/validation set.

The first step of the data pipeline creation is the data preprocessing. At this step we convert the original .csv into the JSON format needed by spaCy. The next step is training a NER model by adding call signs and test words entities. To do this we perform the following steps [6]: create an empty new model, add the new entity label, execute the training procedure, save the trained model, and evaluate the model in order to obtain the quality metrics.

The following outcome has been achieved on the task of call signs NER on the test set. The achieved overall accuracy value is about 0.80. It is obvious that the accuracy is not high enough. Possible solution is increasing the amount of training dataset, including the data from other radio stations. It is a subject of our future research.

References

1. Carreras, Xavier; Màrquez, Lluís; Padró, Lluís (2003). A simple named entity extractor using AdaBoost. CoNLL.

2. Imel, Kathy J.; Hart, James W. (January 2003). "Understanding Wireless Communications in Public Safety". FCC. gov. The National Law Enforcement and Corrections Technology Center–Rocky Mountain. Archived on May 21, 2015. Retrieved September 19, 2015.

3. Kapetanios, Epaminondas; Tatar, Doina; Sacarea, Christian (2013-11-14). Natural Language Processing: Semantic Aspects. CRC Press. p. 298.

4. Jenny Rose Finkel; Trond Grenager; Christopher Manning (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling (PDF). 43rd Annual Meeting of the Association for Computational Linguistics. pp. 363–370.

5. spaCy. [Online]. Available: https://spacy.io/.

6. Custom Named Entity Recognition Using spaCy. [Online]. Available: https://towardsdatascience.com/custom-named-entity-recognition-using-spacy-7140ebbb3718.