










RuREBus: A Case Study of Joint Named Entity Recognition and Relation Extraction from E-Government Domain

Vitaly Ivanin^{1,2}, Ekaterina Artemova³, Tatiana Batura^{4,7},
Vladimir Ivanov^{5,7}, Veronika Sarkisyan³, Elena Tutubalina^{6,7},
and Ivan Smurov^{1,2}^(✉)

¹ ABBYY, Milpitas, USA
ivan.smurov@abbyy.com

² Moscow Institute of Physics and Technology, Dolgoprudny, Russia

³ National Research University Higher School of Economics, Moscow, Russia

⁴ Novosibirsk State University, Novosibirsk, Russia

⁵ Innopolis University, Innopolis, Russia

⁶ Kazan Federal University, Kazan, Russia

⁷ Lomonosov Moscow State University, Moscow, Russia

Abstract. We show-case an application of information extraction methods, such as named entity recognition (NER) and relation extraction (RE) to a novel corpus, consisting of documents, issued by a state agency. The main challenges of this corpus are: 1) the annotation scheme differs greatly from the one used for the general domain corpora, and 2) the documents are written in a language other than English. Unlike expectations, the state-of-the-art transformer-based models show modest performance for both tasks, either when approached sequentially, or in an end-to-end fashion. Our experiments have demonstrated that fine-tuning on a large unlabeled corpora does not automatically yield significant improvement and thus we may conclude that more sophisticated strategies of leveraging unlabelled texts are demanded. In this paper, we describe the whole developed pipeline, starting from text annotation, baseline development, and designing a shared task in hopes of improving the baseline. Eventually, we realize that the current NER and RE technologies are far from being mature and do not overcome so far challenges like ours.

Keywords: Information extraction · Named entity recognition · Relation extraction

1 Introduction

Information extraction tasks, named entity recognition (NER) and relation extraction (RE), have been studied extensively. NER and RE are sometimes

The extended notes for invited talk “When CoNLL-2003 is not Enough: are Academic NER and RE Corpora Well-Suited to Represent Real-World Scenarios?” delivered by Ivan Smurov.