Distant Reading Training School 2020: Named Entity Recognition & Geo-Tagging for Literary Analysis

Ranka Stanković
ranka.stankovic@rgf.bg.ac.rs
University of Belgrade
Faculty of Mining
and Geology
Belgrade, Serbia

PAPER SUBMITTED: 2 November 2021
PAPER ACCEPTED: 15 November 2021

1 Introduction

Distant Reading Training School "Named Entity Recognition & Geo-Tagging for Literary Analysis" was organised virtually within the COST Action 16204: Distant Reading for European Literary History, on 22-25 March 2020. The host institution was the Faculty of Humanities and Social Sciences, University of Rijeka, Croatia, and it was organised within coordinated activity of WG2 "Methods and tools" and WG3 "Literary Theory and History".

WG 2 coordinates activities related to sharing, evaluating and improving methods and tools for distant reading research, with a focus (1) on tool and method adaptation and (2) on establishing best practices across Europe. Members of WG2 come from computational linguistics, text mining, computational stylistics, and digital literary studies. The work of WG3 concentrates on application of distant reading methods to literary history. Members of WG3 come from partners active in digital literary studies and mainstream literary history and theory.

Ten trainers from France, Italy, Norway, Poland, Croatia and Serbia introduced several topics related to Named Entity Recognition (NER) to the target audience, comprising researchers, especially early-career investigators (ECI), from participating countries interested in Distant Reading, Digital Literary Studies, Corpus and Computational Linguistics and/or Literary Theory and their methodological uses across national traditions. The training school included two workshops over the course of 3 days.[1]

[1] All training materials are available on Action's github, including: slides, notebooks and datasets.

2 Workshop 1: Introduction to Named Entity Recognition

The 2-day workshop introduced the task of Named Entity Recognition and described several annotation guidelines and campaigns. The practical part covered a) basic manual annotation with different tools (BRAT,[2] Inception[3] and Recogito),[4] and the analysis of disagreement between annotators, b) automatic annotation with easy-to-use tools such as CLARIN-PL NER tool suite[5] and NER&Beyond,[6] c) TEI-encoding of NER annotation, and d) practical exercises in analysing NE contexts as far as description, sentiment and perception are concerned. For practical reasons and better understanding of the procedures, the exercises were focused on English, but the workshop was addressed to speakers of all ELTeC languages, so that they could learn about NER to work on their collections. Therefore, examples from other languages were also presented.

Diana Santos from the University of Oslo gave an overview of the history of named entity recognition, starting with MUC(K) (1987-1998), IREX (1998), later ACE (2002-2008), CoNLL (2002; 2003), TimeML (2003), ENE (2004), HAREM (2006; 2008), TempEval (2007; 2010; 2013) up to recent SHINRA (2020) and concluding with the references to the research literature about NER.

Carmen Brando from the School for Advanced Studies in the Social Sciences presented several named entity recognition systems through several topics: tool pipelines for linguistic analysis and NER systems; challenges and features for NER systems; types of NER systems; manual annotation and evaluation and training of NER systems; some available out-of-the-box NER systems and output NE annotation formats.

Francesca Frontini from the Institute for Computational Linguistics in Pisa presented the ELTeC NER annotation campaign (Frontini et al. 2020), starting with requirements set by WG3 and motivation for ad hoc annotations, continuing with annotation guidelines containing explanations of category annotations and the annotation procedure (nested annotations,

[2] Brat developer site, github repository: ; Jerteh node used for DR NER 9 language collection

[3] Inception

[4] Recogito

[5] CLARIN-PL

[6] NER&Beyond

inclusion of determiners, ...), and finishing with the process wrap-up and alignment of annotations for various languages.

Ranka Stanković from the University of Belgrade presented the software infrastructure with tools related to NER: BRAT (Stenetorp et al. 2012) for manual annotation and NER&Beyond for formats and transformations. The annotation campaign encompassed the dataset preparation for all languages, dataset publishing (txt+ann), manual annotation or correction of automatic annotations and detailed and simplified annotation. A small experiment with inter-annotator agreement was conducted, to find out where the differences stem from and how to minimise them. Comparison of manual and automatic annotations indicated some problems, testing options and comparison issues (Šandrih Todorović et al. 2021).

Ioana Galleron from the Sorbonne-Nouvelle University and Carmen Brando focused on "translating" the results into TEI[7] (Text Encoding Initiative) annotation, trying to explain what TEI tags for named entities are, how to use them, for simple to more elaborated annotations, and how to convert txt files into TEI/XML file.

Ranka Stanković explained the annotation campaigns (Stanković et al. 2019) and practical work with BRAT, while Maciej Piasecki and Tomasz Walkowiak from the Wrocław University of Technology demonstrated Clarin tools for recognizing named entities and temporal expressions in Polish, English and German.

Carmen Brando gave an introduction to place-based analysis of literary texts: concepts and related work in spatial humanities.

3 Workshop 2: Data Analysis, Representation of the Geo-Entities and Enrichment of the Data Using Wikipedia and Google Maps API

Benedikt Perak from the University of Rijeka introduced Data Analysis, Representation of the Geo-Entities and Enrichment of the Data Using Wikipedia and Google Maps API. Within the *Data analysis task*, using the Google Colab platform and Python scripts, the geo-tagged data was loaded and converted to a Pandas dataframe object as a useful format for creating simple exploratory statistics, e.g. calculating the proportion of the geo-tagged data per language, per book, per period, etc. The *Representation of the*

[7] Text Encoding Initiative

geo-entities comprised two parts: getting the geo-coordinates and mapping geo-names as markers.

Getting the geo-coordinates (longitude and latitude of a place) is a necessary task for the geo-name representation on the map. For finding appropriate geo-data two methods were explored: Google Places API and Wikipedia Python package to explore Wikipedia data on geolocated entities.

The advantage of using the Google Places API[8] to find the geo-coordinates of the geo-name is the possibility to tap into vast information of the Google Places API, which returns information about a variety of categories, places, establishments, prominent points of interest, and geographic locations. One can search for places either by proximity or by text strings, and the Place Search returns a list of places along with summary information about each of them; additional information is available via a Place Details query. The downside of this approach is the need to open an account for this type of query system, which is free for 0–100,000 place requests per month. Using the Wikipedia Python package to explore Wikipedia data on geolocated entities makes accessing and parsing data from Wikipedia easy. The option to find Wikipedia entries by geo-names and geo-coordinates were explored, as well as the extraction of the additional data.

The *Folium Package* was used for mapping geo-names as markers and representing the data as markers with tooltip and HTML popup on the map. This representation helps literary scholars with the location of the narratives and the interpretation of literary texts.

The teaching materials included several Colab notebooks: NER Processing using NLP tools (spaCy), Data Analysis, Representation of the Geo-Entities and Enrichment of the Data Using Wikipedia and Python Client for Google Maps Services. The Jupyter Notebooks are available in the WG2_notebooks folder.

Acknowledgment

We are grateful to Antonija Primorac, a Working Group 3 Leader and Joanna Byszuk, a Working Group 2 Leader for organizing this school and to all the trainers for their valuables lectures.

[8] Places API

References

Frontini, Francesca, Carmen Brando, Joanna Byszuk, Ioana Galleron, Diana Santos, and Ranka Stanković. 2020. “Named entity recognition for distant reading in ELTeC.” In *CLARIN Annual Conference 2020.*

Šandrih Todorović, Branislava, Cvetana Krstev, Ranka Stanković, and Milica Ikonić Nešić. 2021. “Serbian NER& Beyond: The Archaic and the Modern Intertwined.” In *Deep Learning Natural Language Processing Methods and Applications – Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2021),* edited by Galia Angelova, Maria Kunilovskaya, Ruslan Mitkov, and Ivelina Nikolova-Koleva, 1252–1260. INCOMA Ltd. ISBN: 978-954-452-072-4. https://doi.org/10.26615/978-954-452-072-4_141.

Stanković, Ranka, Diana Santos, Francesca Frontini, Tomaž Erjavec, and Carmen Brando. 2019. “Named Entity Recognition for Distant Reading in Several European Literatures.” *DH Budapest 2019.*

Stenetorp, Pontus, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun’ichi Tsujii. 2012. “BRAT: a web-based tool for NLP-assisted text annotation.” In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics,* 102–107.