Using deep learning to explore movement of people in a large corpus of biographies

The Austrian Biographical Dictionary (ÖBL) offers massive amounts of data for historical research (we are currently working on exploring how well this resource is actually suited for research). It contains well over 18.000 biographies of important people who died between 1815 and 1950. While this resource has a lot of (scientific) shortcomings, it is probably the only dataset that gives an overview of the Austrian-Hungarian Empire and succeeding states (the first and second republic) from the peoples perspectives.

Other projects have already successfully shown that automatically structuring biographical data can give researchers an overview of societal structures1 that cant be achieved with non-digital means. In our article we will demonstrate a hybrid approach that we designed to semantically enrich the ÖBL.

We have developed a web application that researchers can use to semantically annotate subsets of the ÖBL by hand2. The tool offers direct access to reference resources - such as geonames3 or the "Gemeinsame Normdatei”4 - and makes it an easy task to annotate semantic relations between portrayed persons and other entities (places, persons, institutions etc.). We use these manually created annotations to (re)train Named Entity Recognizers (NER) on the one hand and deep learning models that predict the kind of relation between the named entities (e.g. a place) and the portrayed person on the other. These deep learning models use the nature of biographies to simplify the classification task: in almost all cases the subject of a sentence is the portrayed person. Our relation extraction pipeline therefore does the following:

- Search for named entities in the sentences (in our case places)
- Parse the sentence; from the named entity run up the parse tree and extract all tokens that fulfill certain part-of-speech (POS) tag rules (e.g. POS tag == verb).
- Convert this list of tokens into an array of token ids, feed the model with this array and predict the probabilities of the various labels.

In our contribution we will use the movement of people5 within the ÖBL corpus as an example to explore the capacities of these models (we trained several models using different features). The models trained from manually tagged biographies are not only evaluated against a gold standard, but also against a rule based approach realized in GATE6. Our contribution discusses the advantages and downsides of both approaches - deep learning models and rule based pipelines - with regard to:

- effort to create/implement
- accuracy
- and ease of improving accuracy/removing errors

While the main goal of our contribution is to showcase the technologies and compare them we will also give first insights into migration/movement patterns in the Austrian-Hungarian Empire7 between 1815 and 1950. The analysis focusses on important tipping points of the Austrian History: 1814/15, 1848, 1867, 1914, 1918/19, 1934, 1938, 1945, 1955 and compares the migration/movement patterns of these years.

1. CN Warren, D Shore, J Otis, and L Wang, “Six Degrees of Francis Bacon: A Statistical Method for Reconstructing Large Historical Social Networks.,” DHQ: Digital Humanities Quaterly, 2016. ↩

[2] Please see https://apisdev.acdh.oeaw.ac.at for a demo version of the tool ↩

[3] http://www.geonames.org/ ↩

[4] http://www.dnb.de/DE/Standardisierung/GND/gnd_node.html ↩

[5] This includes migrations, journeys, long stays abroad and others more. ↩

[6] https://gate.ac.uk/ ↩

[7] and its predecessor and successor states ↩