NERDPool

Datenpool für Named Entity Recognition

Andorfer, Peter
peter.andorfer@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Austria

Schlögl, Matthias
matthias.schloegl@oeaw.ac.at
Österreichische Akademie der Wissenschaften, Austria

Bleier, Roman
roman.bleier@uni-graz.at
Universität Graz, Austria

Bedeutung

In digitalen Editionen ist die automatische Erkennung und Annotation von Personen, Orten und Datumsangaben eine wichtige Aufgabe, die langfristig die händische Annotation ablösen wird. Machine Learning (ML) und Named Entity Recognition (NER) spielt dabei eine zentrale Rolle.[1] Historische Texte bilden noch ein Problem, da oft zu wenig Trainingsmaterial zur Verfügung steht, um entsprechende ML-Modelle zu trainieren. Andererseits werden seit über 30 Jahren digitale Editionen mit strukturierten Daten produziert, die diese Lücke füllen könnten.

Das von CLARIAH-AT finanzierte Projekt NERDPool versucht einerseits existierende (XML/TEI kodierte) Editionsdaten zu nutzen und daraus einen Pool an Trainingsdaten zu generieren, sowie andererseits Workflows zu erproben und zu implementieren, die es erlauben, einfach und effizient bestehende Korpa manuell zu annotieren. Den Schwerpunkt setzt das Projekts auf frühneuzeitliche deutsche Texte, ein Sprachstufe für die es wenig NER Material gibt. Die Datensätze werden über die Webapplikation https://nerdpool-api.acdh-dev.oeaw.ac.at/ respektive über eine impelementierte offene API veröffentlicht und können etwa mit Hilfe eines eigenen Python-Clients (https://github.com/acdh-oeaw/nerdpool-client, 14. Juli 2021) heruntergeladen werden. Mit Stand Juli umfasst NERDPool rund 23.500 annotierte Datensätze. Darunter sind etwa Akten vom Regensburger Reichstag von 1576 (*https://reichstagsakten-1576.uni-graz.at*), Ministerratsprotokolle Österreichs und der österreichisch-ungarischen Monarchie 1848–1918 oder die ersten Ausgaben des Wienerischen Diariums (um 1750).

XML/TEI → Annotationen

Die in NERDPool gesammelten Daten sind stets das Resultat manueller Annotation. Die konkrete Annotationsarbeit erfolgte im Kontext der Erstellung einer XML/TEI kodierten Digitalen Edition. Hier wurden Personen, Orte, Datumsangaben mit entsprechenden TEI Tags annotiert. Die Daten werden über die GitHub API direkt von einem Repo abgerufen und dahingehend weiterverarbeit, als die annotierten Textknoten gelesen und die Offsets der annotierten Elemente extrahiert werden (https://github.com/acdh-oeaw/acdh-tei-pyutils). Konkret wird ein Nodeset wie `<p><placeName>Wien</placeName> ist eine Stadt.</p>` in folgenden JSON-Eintrag `{“text”: “Wien ist eine Stadt.”, “entities”: [0, 3, “LOC”]}` konvertiert und anschließend in die Django basierte Webapplikation nerdpool-api importiert.

Prodigy & „custom loaders“

Ein zweiter Ansatz setzt auf das Annotationstoolkit Prodigy (https://prodi.gy/). Das kostenpflichtige und teilweise closed sourced Softwarepaket bietet ein äußerst effizientes Annotationsinterface und lässt sich sehr gut adaptieren, beispielsweise durch das Hinzufügen sogenannter ‘custom loaders’, welche Textdaten in das Annotationsinterface streamen und es so etwa erlauben Texte aus bestehenden APIs mit Prodigy zu annotieren. Mit einem solchen Loader „pr_transkribus.py“ (https://github.com/acdh-oeaw/acdh-prodigy-utils/blob/master/pr_transkribus.py) wurden etwa Texte direkt aus Transkribus über die Transkribus-API (https://transkribus.eu/TrpServer/Swadl/wadl.html) in Prodigy geladen.

Die Orchestrierung der einzelnen Prodigy-Instanzen, das notwendige Usermanagement der einzelnen Annotator*Innen sowie die Sicherung und Zusammenführung der Annotationsdaten erfolgt mittels Django, Postgresql, Nginx und dem Container Management Tool Portainer (ptr target="https://github.com/acdh-oeaw/nerdpool"/>).

Probleme und Lösungen

In der konkreten Implementierung der obene beschrieben Workflows bereitete vor allem die für Prodigy notwendige Tokenisierung und die darauf aufbauende Segmentierung (Sentence-Splitting) Probleme:

Die Syntax und Satzlänge historischer weicht teils massiv von jenen zeitgenössischer Texte - welche gemeinhin zum Training von NLP Modellen verwendet werden - ab.

In historischen Texten finden sich viele zum Teil heute nicht mehr gängige Abkürzungen (https://abbr.acdh.oeaw.ac.at) bzw. Trenn- und Satzzeichen - was sich wiederum ungünstig auf die Tokenisierung auswirkt.

Was das Problem einer automatisierten Satzsegmentierung betrifft, so wurde darauf zum Teil verzichtet und die Texte anhand von formalen Kriterien wie beispielsweise manuell annotierter (XML/TEI) oder von Layouterkennung erkannter Absätze geteilt. Dies hat den Vorteil, dass die Annotationssamples nicht an den falschen Stellen unterbrochen werden, führt aber Teilweise zu sehr langen Annotationssamples, welche das Annotieren vor allem über ein auf Effizienz ausgerichtete System wie Prodigy erschwert.

In einem anderen Ansatz wurde das zur Tokenisierung verwendete Spacy Modell um eine Liste von Abkürzungen erweitert. Das funktioniert tendenziell gut, bringt allerdings einen (weiteren) technisch-administraiven Overhead hinischtlich der Verwaltung der (Tokenisierungs-)Modelle mit sich.

Protokolle des österreichischen Ministerrates als Beispiel

Das Potential der gesammelten Annotationsdaten soll beispielhaft an der bereits erwähnten Edition der "Protokolle des österreichischen Ministerrates 1848-1867" (MRP) gezeigt werden. Auf Basis von rund 12.000 manuell annotierten Samples wurde ein spaCy NER Modell (Version 3.x) trainiert (ptr target="https://huggingface.co/csae8092/de_MRP_NER"/>). Während das kleine spaCy Standardmodell für Deutsch auf dem Evaluationsset der MPR Daten F1 Werte für Personen und Organisationen von rund 23 bzw. 12 Prozent erzielt, liegen die Werte beim MRP Modell bei 91 und 82 Prozent. Die Werte für die weiters annotierten Kategorien LOC und GPE liegen bei 87 bzw. 58 Prozent (ptr target="https://github.com/csae8092/ner-tei-playgrounds"/>). Das MRP-Modell ist damit trotz historischer Sprachstufe, vielen Abkürzungen und vergleichsweise wenigen Trainingsdaten nur knapp unter aktuellen Named Entity Recognizern.

Fußnoten

1. Vgl. dazu die in der Bibliographie angeführte Literatur.

Bibliographie

Urbano, J et al. (2012): “Named Entity Recognition: Fallacies, challenges and opportunities”, *Computer Standards & Interfaces* 35(5): pp. 482–489. doi: 10.1016/j.csi.2012.09.004 [letzter Zugriff 13. Juli 2021].

Kettunen, Kimmo / Mäkelä, Eetu / Ruokolainen, Teemu / Kuokkala, Juha / Löfberg, Laura (2017): “Old Content and Modern Tools: Searching Named Entities in a Finnish OCRed Historical Newspaper Collection 1771–1910”, in: *Digital Humanities Quarterly* 11(3), http://digitalhumanities.org/dhq/vol/11/3/000333/000333.html [letzter Zugriff 13. Juli 2021].

Kannisto, Maiju / Kauppinen, Pekka (2020): "Of Great Men and Eurovision Songs: Studying the Finnish Audio-Visual Heritage through NER-based Analysis on Metadata", in: Fridlund, Mats / Oiva, Mila / Paju, Petri (eds.) *Digital Histories: Emergent Approaches within the New Digital History*, 165-180.