Zurück in die Zukunft: Named-Entity-Recognition für digitale (historisch-kritische) Editionen

Tamara Haeder
Hochschule Darmstadt
Fachbereich Mathematik & Naturwissenschaften und Fachbereich Informatik

Zusammenfassung

Die Mehrheit der Natural Language Processing Untersuchungen im Rahmen der Digital Humanities betonen die Schwierigkeiten, welche sich durch diachronische Daten ergeben. Dabei können andere Bereiche bedeutende Verluste der Performanz von Natural Language Processing Verfahren verursachen. Verfahren, die Sequenzlabelling Probleme modellieren, basieren oft auf Wahrscheinlichkeitsverteilungen der Worte und ihrer Label. Wenn diese nicht konsequent konsistent annotiert werden, können zum Beispiel Named-Entity-Recognition Verfahren keine eindeutige Zuordnung der Worte zu ihren korrespondierenden Entitäten vornehmen. Neben einem einheitlichen Annotationsschema innerhalb einer digitalen (historisch-kritischen) Edition, ist die Etablierung von allgemeinen Standards wichtig, damit adaptierbare Modelle generiert werden können.

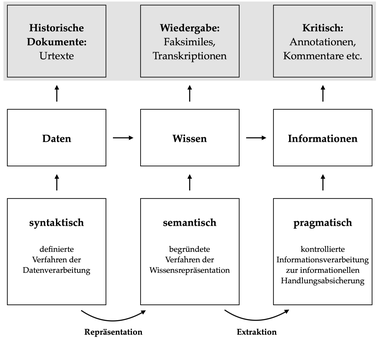
Einleitung

Digitale historisch-kritische Editionen werden immer populärer. Vor langer Zeit hätte die Bezeichnung noch für Verwunderung gesorgt, da zeitens die Verwendung von ”digitalünd ”historischïn einem Atemzug als Oxymoron, als zwei sich ausschließende Begriffe, galt. Mit zunehmender Digitalisierung wurde auch das Potential dieser in den Literatur- und Gesellschaftswissenschaften erkannt. Es entstand eine neue Disziplin, die der Digital Humanities, wodurch die Vereinbarkeit digitaler und historischer Konzepte geschaffen wurde.

Editor*innen sind maßgebend für die Qualität der Editionen und bilden die Grundlage für die Transformation von einer digitalisierten Druckedition in eine digitale (historisch-kritische) Edition. Erst durch ihre kritische Auseinandersetzung mit den Dokumenten wird eine digitale Edition zu einer digitalen historisch-kritischen Edition. Sie schlagen die Brücke zwischen Vergangenheit und Gegenwart.

Um die digitalen (historisch-kritischen) Editionen der Zukunft aufzubauen, werden Computerwissenschaftler*innen im Aufbau hinzugezogen. Durch Machine Learning Verfahren, insbesondere im Bereich Data Mining und Natural Language Processing, kann der Arbeitsprozess von Editor*innen effizienter gestaltet werden.

Aus der Sicht von Computerwissenschaftler*innen stellen digitale (historisch-kritische) Editionen Information Retrieval Systeme dar. Information Retrieval (IR) beschreibt den Prozess der Generation von Informationen aus Daten. Historisch-kritische Editionen generieren Informationen in Form von Kommentaren und Annotationen aus den Daten, nämlich den Urdokumenten.



Digitale (historisch-kritische) Editionen als IR-System (Quelle:[1], S.20). Der grau hinterlegte Abschnitt, stellt eine Erweiterung dieser Graphik dar.

Innerhalb der im Rahmen dieser Arbeit durchgeführten Umfrage Ëffizientes Suchen und Editieren von Dokumenten für Nutzer*innen und Editor*innen digitaler (historisch-kritischer) Editionen durch Natural Language Processing”hat sich abgezeichnet, dass Editor*innen vor allem an Named-Entity-Recognition Tools interessiert sind, die bei der Annotation der Dokumente unterstützt.

Named-Entity-Recognition

Named Entity Recognition (auch: Entity Chunking, Entity Extraction oder Entity Identification) ist ein Teilgebiet der Information Extraction (IE). NERAlgorithmen erkennen und benennen vordefinierte Informationseinheiten (sog. Entitäten). Je nach Anwendungsfall gibt es unterschiedliche Auslegungen, welche Einheiten und auf welchem Granularitätniveau, Einheiten als Entitäten definiert werden. Es gibt jedoch konventionelle Entitäten, welche unumstritten als solche anerkannt sind: Person, Ort, Organisationen, Datum und Geld.

Conditional Random Fields

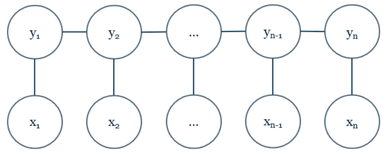


Diagramm Conditional Random Field

Named-Entity-Recognition wird als Sequenzlabeling Problem modelliert, sodass gegeben einer beobachteten Inputsequenz X, jedem Element der Sequenz ein Label Y zugewiesen wird. Im Folgenden seien $x = (x_1, \ldots, x_n)$ und $y = (y_1, \ldots, y_n)$ Realisierungen von X und Y.

Conditional Random Fields modellieren die Verteilung der Label wie folgt:[**?**]

$$P(Y|X) = \prod_{i=1}^{n} P(y_i|y_{i-1}, x) \tag{1}$$

Zusätzlich können neue Feature definiert werden, die durch die Featurefunktion f ausgedrückt werden und mit w gewichtet werden.

$$P(y_i|y_{i-1}, x) = \frac{1}{Z(x)} \exp(\sum_{l=1}^{J} \sum_{i=1}^{n} w_l f_l(y_{i-1}, y_i, x, i)), \tag{2}$$

wobei der Nenner $Z(x)$ zur Normalisierung über alle möglichen Labelsequenzen dient:

$$Z(x) = \sum_{y} \exp(\sum_{l=1}^{J} \sum_{i=1}^{n} w_l f_l(y_{i-1}, y_i, x, i)) \tag{3}$$

Die optimale Labelsequenz ist wie folgt definiert:

$$\hat{Y} = {}_Y \; P(Y|X) \tag{4}$$

$$= {}_Y \sum_{l=1}^{K} \sum_{i=1}^{n} w_i f_i(y_{i-1}, y_i, x, i) \tag{5}$$

Frank Wedekind Briefedition

Die Arbeit wird im Rahmen des Projekts Ëdition der Korrespondenz Frank Wedekinds als Online-Volltextdatenbank”der Hochschule Darmstadt und Universität Mainz durchgeführt. Der praktische Teil der Arbeit implementiert ein Named-Entity-Recognition System basierend auf einem Conditional Random Field für die Briefe von und an Frank Wedekind. Es liegen 1150 annotierte Dokumente vor, welche folgende Entitäten beinhalten:

- Person
- Ort
- Örtlichkeit
- Ereignis
- Werk

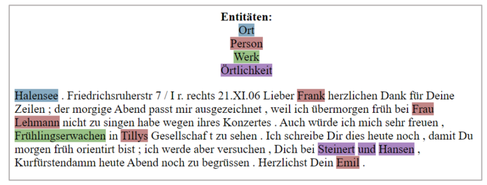
Entitätsreferenzen und die korrespondierenden Entitäten werden aus den mit HTML annotierten Briefen mittels Regex-Expressions extrahiert. Zusätzlich werden noch bestehende Kommentare der Editor*innen entfernt. Diese würden einerseits Schwierigkeiten beim Chunking verursachen und andererseits entsprechen sie nicht der Sprache von den Autor*innen der Briefe.

Die bereinigten Dokumente werden tokeniziert und jeder Token wird mit der entsprechenden Entität gelabelt. Hierbei werden zwei unterschiedliche Schemata verfolgt. Einmal werden die Token so gelabelt, dass sie dem Annotationsschema der Editor*innen folgen, welche bei Mehrfachnennungen derselben Entitätsreferenz immer nur das erste Aufkommen annotieren. Zum Anderen wird ein Datensatz erstellt, der einem konsistenten Schema folgt und jede Entitätsreferenz annotiert.

Während des Data Engineering Prozesses werden weitere relevante Feature generiert. Darunter zählen u.a. der Part-of-Speech-Tag, das Lemma und Präfix und Suffix eines Token.

Es werden fünf verschiedene Modelle miteinander verglichen und entsprechend ihrer Precision, ihres Recalls und des F_1-Scores evaluiert. Die Ergebnisse zeigen, dass Modelle, die mit einem konsistenten Annotationsschema trainiert werden besser performen als diejenigen, die innerhalb eines Briefes nur die erste Entitätsreferenz annotieren. Personen, Orte und Örtlichkeiten werden vom System am ehesten richtig annotiert. Die restlichen Entitäten sind wohl durch ihre geringe Fallzahl schwer zu trainieren.

Da das Modell, welches auf einem konsistenten Datensatz trainiert wurde am besten performt, wird dieses für die Frank Wedekind Briefedition implementiert. Um dem Annotationsschema der Editor*innen zu entsprechen, werden die Vorhersagewerte so aufbereitet, dass bei Mehrfachreferenzen die erste Referenz die erkannte Entität erhält und alle anderen Referenzen auf Öther”gesetzt werden. Dieses Vorgehen erhöht die Performanz, sodass das finale Modell eine Precision von 0.677, einen Recall von 0.556 und einen F_1-Score von 0.605 erzielt.



Ergebnis eines Beispieldokumentes

Fazit und Ausblick

Conditional Random Fields performen besser auf Daten mit einem konsistenten Annotationsschema. Eine andere und möglicherweise besserer Ansatz für die zugrundeliegenden Daten könnte ein BiLSTM Neuronales Netz darstellen, da sie durch ihre Struktur einen ganzen Brief und nicht nur einzelne Sätze in Betracht ziehen können.

Um Editor*innen noch weiter in ihrem Editionsprozess unterstützen zu können und eine digitale (historisch-kritische) Edition mit weiteren Funktionalitäten auszustatten, sind weitere Untersuchungen notwendig, welche den Bedarf nach solchen erheben. Named-Entity-Recognition setzt den Grundstein für weitere Tools wie zum Beispiel Relation Extraction.

Literatur

[1] Andreas Henrich. Information retrieval 1 (grundlagen, modelle und anwendungen), 2008.

[2] Bengong Yu and Zhaodi Fan. A comprehensive review of conditional random fields: variants, hybrids and applications. *Artificial Intelligence Review*, pages 1–45, 2019.