Maschinelles Lernen lernen: Ein CRETA-Hackatorial zur reflektierten automatischen Textanalyse

Kerstin Jung, Gerhard Kremer

Einleitung

Das Ziel dieses Tutorials ist es, den Teilnehmerinnen und Teilnehmern konkrete und praktische Einblicke in einen Standardfall automatischer Textanalyse zu geben. Am Beispiel der automatischen Erkennung von Entitätenreferenzen gehen wir auf allgemeine Annahmen, Verfahrensweisen und methodische Standards bei maschinellen Lernverfahren ein. Die Teilnehmerinnen und Teilnehmer können beim Bearbeiten von lauffähigem Programmiercode den Entscheidungsraum solcher Verfahren ausleuchten und austesten. Es werden dabei keinerlei Vorkenntnisse zu maschinellem Lernen oder Programmierkenntnisse vorausgesetzt.

Es gibt keinen Grund, den Ergebnissen von maschinellen Lernverfahren im Allgemeinen und NLP-Tools im Besonderen blind zu vertrauen. Durch die konkreten Einblicke in den "Maschinenraum" von maschinellen Lernverfahren wird den Teilnehmenden ermöglicht, das Potenzial und die Grenzen statistischer Textanalysewerkzeuge realistischer einzuschätzen. Mittelfristig hoffen wir dadurch, den immer wieder auftretenden Frustrationen beim Einsatz automatischer Verfahren für die Textanalyse und deren teilweise wenig zufriedenstellender Ergebnis-Daten zu begegnen, aber auch die Nutzung und Interpretation der Ergebnisse von maschinellen Lernverfahren (d.h. in erster Linie von automatisch erzeugten Annotationen) zu fördern. Zu deren adäquater Nutzung, etwa in hermeneutischen Interpretationsschritten, ist der Einblick in die Funktionsweise der maschinellen Methoden unerlässlich. Insbesondere ist die Art und Herkunft der Trainingsdaten für die Qualität der maschinell produzierten Daten von Bedeutung, wie wir im Tutorial deutlich machen werden.

Neben einem Python-Programm für die automatische Annotierung von Entitätenreferenzen, mit und an dem während des Tutorials gearbeitet werden wird, stellen wir ein heterogenes, manuell annotiertes Korpus sowie die Routinen zur Evaluation und zum Vergleich von Annotationen zu Verfügung. Das Korpus enthält Entitätenreferenzen, die im "Center for Reflected Text Analytics" (CRETA)[1] in den letzten zwei Jahren annotiert wurden, und deckt Texte verschiedener Disziplinen und Sprachstufen ab.

[1] www.creta.uni-stuttgart.de

Entitätenreferenzen

Als empirisches Phänomen befassen wir uns mit dem Konzept der Entität und ihrer Referenz. Das Konzept steht für verschiedene linguistische und semantische Kategorien, die im Rahmen der Digital Humanities von Interesse sind. Es ist bewusst weit gefasst und damit anschlussfähig für verschiedene Forschungsfragen aus den geistes- und sozialwissenschaftlichen Disziplinen. Auf diese Weise können unterschiedliche Perspektiven auf Entitäten berücksichtigt werden. Insgesamt werden in den ausgewählten Texten fünf verschiedene Entitätenklassen betrachtet: PER (Personen/Figuren), LOC (Orte), ORG (Organisationen), EVT (Ereignisse) und WRK (Werke).

Unter Entitätenreferenzen verstehen wir Ausdrücke, die auf eine Entität in der realen oder fiktiven Welt referieren. Das sind zum einen Eigennamen (Named Entities, z.B. "Peter"), zum anderen Gattungsnamen (z.B. "der Bauer"), sofern diese sich auf eine konkrete Instanz der Gattung beziehen. Dabei wird als Referenzausdruck immer die maximale Nominalphrase (inkl. Artikel, Attribut) annotiert. Pronominale Entitätenreferenzen werden hingegen nicht annotiert.

In **literarischen Texten** sind vor allem Figuren und Räume als grundlegende Kategorien der erzählten Welt von Interesse. Über die Annotation von Figurenreferenzen können u.a. Figurenkonstellationen und -relationen betrachtbar gemacht sowie Fragen zur Figurencharakterisierung oder Handlungsstruktur angeschlossen werden. Spätestens seit dem *spatial turn* rückt auch der Raum als relevante Entität der erzählten Welt in den Fokus. Als "semantischer Raum" (Lotmann, 1972) übernimmt er eine strukturierende Funktion und steht in Wechselwirkung mit Aspekten der Figur.

In den **Sozialwissenschaften** sind politische Parteien und internationale Organisationen seit jeher zentrale Analyseobjekte der empirischen Sozialforschung. Die Annotation der Entitäten der Klassen ORG, PER und LOC in größeren Textkorpora ermöglicht vielfältige Anschlussuntersuchungen, unter anderem zur Sichtbarkeit oder Bewertung bestimmter Instanzen, beispielsweise der Europäischen Union.

Textkorpus

Die Grundlage für (überwachte) maschinelle Lernverfahren bilden Annotationen. Um die Annotierung von Entitätenreferenzen automatisieren zu können, bedarf es Textdaten, die die Vielfalt des Entitätenkonzepts abdecken. Bei diesem Tutorial werden wir auf Annotationen zurückgreifen, die im Rahmen von CRETA an der Universität Stuttgart entstanden sind (cf. Blessing et al., 2017; Reiter et al., 2017a). Das Korpus enthält literarische Texte aus zwei Sprachstufen des Deutschen (Neuhochdeutsch und Mittelhochdeutsch) sowie ein

sozialwissenschaftliches Teilkorpus. [2]

Der ***Parzival*** **Wolframs von Eschenbach** ist ein arthurischer Gralroman in mittelhochdeutscher Sprache, entstanden zwischen 1200 und 1210. Der *Parzival* zeichnet sich u.a. durch sein enormes Figureninventar und seine komplexen genealogischen Strukturen aus, wodurch er für Analysen zu Figurenrelationen von besonderem Interesse ist. Der Text ist in 16 Bücher unterteilt und umfasst knapp 25.0000 Verse.

Johann Wolfgang von Goethes *Die Leiden des jungen Werthers* ist ein Briefroman aus dem Jahr 1774. Unsere Annotationen sind an einer überarbeiteten Fassung von 1787 vorgenommen und umfassen die einleitenden Worte des fiktiven Herausgebers sowie die ersten Briefe von Werther an seinen Freund Wilhelm.

Das **Plenardebattenkorpus des deutschen Bundestages** besteht aus den von Stenografinnen und Stenografen protokollierten Plenardebatten des Bundestages und umfasst 1.226 Sitzungen zwischen 1996 und 2015.[3] Unsere Annotationen beschränken sich auf Auszüge aus insgesamt vier Plenarprotokollen, die inhaltlich Debatten über die Europäische Union behandeln. Hierbei wurde pro Protokoll jeweils die gesamte Rede eines Politikers bzw. einer Politikerin annotiert.

Ablauf

Der Ablauf des Tutorials orientiert sich an sog. *shared tasks* aus der Computerlinguistik, wobei der Aspekt des Wettbewerbs im Tutorial vor allem spielerischen Charakter hat. Bei einem traditionellen *shared task* arbeiten die teilnehmenden Teams, oft auf Basis gleicher Daten, an Lösungen für eine einzelne gestellte Aufgabe. Solch eine definierte Aufgabe kann z.B. *part of speech-tagging* sein. Durch eine zeitgleiche Evaluation auf demselben Goldstandard können die entwickelten Systeme direkt verglichen werden. In unserem Tutorial setzen wir dieses Konzept live und vor Ort um.

Zunächst diskutieren wir kurz die zugrundeliegenden Texte und deren Annotierung. Annotationsrichtlinien werden den Teilnehmerinnen und Teilnehmern im Vorfeld zur Verfügung gestellt. Im Rahmen der Einführung wird auch auf die konkrete Organisation der Annotationsarbeit eingegangen, so dass das Tutorial als Blaupause für zukünftige Tätigkeiten der Teilnehmenden in diesem und ähnlichen Arbeitsfeldern dienen kann.

Die Teilnehmerinnen und Teilnehmer versuchen selbständig und unabhängig voneinander, eine Kombination aus maschinellen Lernverfahren, Merkmalsmenge und Parametersetzungen zu finden, die auf einem neuen, vom automatischen

[2] Aus urheberrechtlichen Gründen wird das Tutorial ohne das Teilkorpus zu Adornos ästhetischer Theorie stattfinden, das in den Publikationen erwähnt wird.

[3] Die Texte wurden im Rahmen des PolMine-Projekts verfügbar gemacht: http://polmine.sowi.uni-due.de/polmine/

Lernverfahren ungesehenen Datensatz zu den Ergebnissen führt, die dem Goldstandard der manuellen Annotation am Ähnlichsten sind. Das bedeutet konkret, dass der Einfluss von berücksichtigten Features (z.B. Groß- und Kleinschreibung oder Wortlänge) auf die Erkennung von Entitätenreferenzen empirisch getestet werden kann. Dabei sind Intuitionen über die Daten und das annotierte Phänomen hilfreich, da simplem Durchprobieren aller möglichen Kombinationen ("brute force") zeitlich Grenzen gesetzt sind.

Wir verzichten bewusst auf eine graphische Benutzerschnittstelle (cf. Reiter et al., 2017b) – stattdessen editieren die Teilnehmerinnen und Teilnehmer das (Python)-Programm direkt, nach einer Einführung und unter Anleitung. Vorkenntnisse in Python sind dabei nicht nötig: Das von uns zur Verfügung gestellte Programm ist so aufgebaut, dass auch Python-Neulinge relativ schnell die zu bearbeitenden Teile davon verstehen und damit experimentieren können. Wer bereits Erfahrung im Python-Programmieren hat, kann fortgeschrittene Funktionalitäten des Programms verwenden.

Wie am Ende jedes maschinellen Lernprozesses wird auch bei uns abschließend eine Evaluation der automatisch generierten Annotationen durchgeführt. Hierfür werden den Teilnehmerinnen und Teilnehmern nach Ablauf einer begrenzten Zeit des Experimentierens und Testens (etwa 60 Minuten) die finalen, vorher unbekannten Testdaten zur Verfügung gestellt. Auf diese Daten werden die erstellten Modelle angewendet, um automatisch Annotationen zu erzeugen. Diese wiederum werden dann mit dem Goldstandard verglichen, wobei die verschiedenen Entitätenklassen sowie Teilkorpora getrennt evaluiert werden. Auch das Programm zur Evaluation stellen wir bereit.

Lernziele

Am hier verwendeten Beispiel der automatischen Annotation von Entitätenreferenzen demonstrieren wir, welche Schritte für die Automatisierung einer Textanalyseaufgabe mittels maschinellen Lernverfahren nötig sind und wie diese konkret implementiert werden können. Die Teilnehmerinnen und Teilnehmer bekommen einen zusammenhängenden Überblick von der manuellen Annotation ausgewählter Texte über die Feinjustierung der Lernverfahren bis zur Evaluation der Ergebnisse. Die vorgestellte Vorgehensweise für den gesamten Ablauf ist grundsätzlich auf ähnliche Projekte übertragbar.

Das Tutorial schärft dabei das Verständnis für den Zusammenhang zwischen untersuchtem Konzept und den dafür relevanten Features, die in ein statistisches Lernverfahren einfließen. Durch Einblick in die technische Umsetzung bekommen die Teilnehmerinnen und Teilnehmer ein Verständnis für die Grenzen und Möglichkeiten der Automatisierung, das sie dazu befähigt, zum einen das Potenzial solcher Verfahren für eigene Vorhaben realistisch(er) einschätzen zu können, zum anderen aber auch Ergebnisse, die auf Basis solcher Verfahren erzielt wurden, angemessen hinterfragen und deuten zu können.

Referenzen

Kuhn, Jonas / Reiter, Nils (2015): "A Plea for a Method-Driven Agenda in the Digital Humanities" in: *Digital Humanities 2015: Conference Abstracts*, Sydney.

Reiter, Nils / Blessing, Andre / Echelmeyer, Nora / Koch, Steffen / Kremer, Gerhard / Murr, Sandra / Overbeck, Maximilian / Pichler, Axel (2017a): "CUTE: CRETA Unshared Task zu Entitätenreferenzen" in *Konferenzabstracts DHd2017*, Bern.

Reiter, Nils / Kuhn, Jonas / Willand, Marcus (2017b): "To GUI or not to GUI?" in *Proceedings of INFORMATIK 2017*, Chemnitz.

Blessing, Andre / Echelmeyer, Nora / John, Markus / Reiter, Nils (2017): "An end-to-end environment for research question-driven entity extraction and network analysis" in *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, Vancouver.

Lotman, Juri (1972): *Die Struktur literarischer Texte*, München.