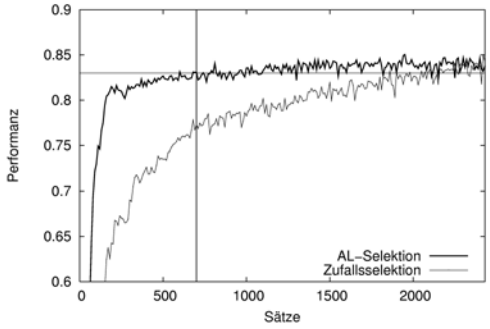
Automatische Erkennung und effiziente Annotation von anonymisierungsrelevanten Begriffen in klinischen Freitexten

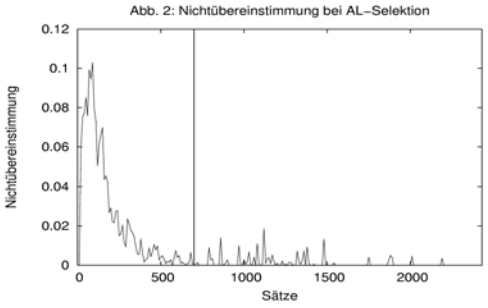
Wermter J, Tomanek K, Balzer F
Jena University Language and Information Engineering (JULIE) Lab, Friedrich-Schiller-Universität Jena, Deutschland
joachim.wermter@uni-jena.de

Einleitung und Fragestellung. Patienten wird im modernen Krankenhausbetrieb eine vertrauliche Behandlung aller sensiblen Daten, die sich im Laufe einer Behandlung ansammeln, zugesichert. Das Interesse an diesen Daten ist groß: Nicht nur die behandelnden Ärzte und andere Fachkräfte können sie für ihre Arbeit nutzen, sondern auch in der Forschung und Lehre besteht ein Bedarf, da Fallbeschreibungen aus der Praxis einen großen Fundus an Informationen darstellen. Darüber hinaus hat die statistische Auswertung von klinik- und krankenhausinternen medizinischen Dokumenten einen großen Stellenwert: die Masse an aufbereiteten Daten, die durch Methoden zur Anonymisierung bzw. Pseudonymisierung zur Verfügung stünde, würde es erlauben, gezielten Fragestellungen nachzugehen (wie z.B. der Wirksamkeit von Operationsverfahren, der Erstellung onkologischer Statistiken, etc.) und eine bessere Selektion von potentiellen Teilnehmern an klinischen Studien wäre möglich. Die dem Patienten zugesicherte Vertraulichkeit ist jedoch schwer zu wahren. Zum Einen ist es sehr langwierig, Daten manuell zu anonymisieren, schließt aber zum Anderen trotzdem nicht die Gefahr aus, dass sich Rückschlüsse auf die Identität des Patienten ziehen lassen. So wurden in einem Experiment bis zu 6 % der sensiblen Informationen von einem Menschen nicht als solche markiert und entsprechend anonymisiert [1].
Daher wäre eine weitgehend automatisierte Erkennung und Anonymisierung medizinischer Daten wünschenswert, wobei in diesem Zusammenhang vor allem die Arbeit von Sweeney [1] zu erwähnen ist. Das von Sweeney realisierte „Scrub System“ identifiziert in englischen Klinik-Dokumenten Zeichenketten, die Rückschlüsse auf die Identität von Patienten erlauben. Durch die Anwendung von Methoden, die nach der Struktur von persönlichen Informationen suchen (z.B. Identitätsmarker wie „Herr/Frau“ für Personennamen), und durch die Kenntnis häufig vorkommender Namen sowie Daten aus den Krankenakten konnte eine Trefferquote von 99-100% erzielt werden. Typischerweise befinden sich eine Vielzahl der anonymisierungsrelevanten Begriffe (wie z.B. Name, Geburtsdatum und Adresse des Patienten, sowie behandelnder Arzt, Abteilungsnamen) in den strukturierten Abschnitten medizinischer Dokumenten: In den markierten Dateiköpfen von Arztbriefen, aber auch in den Stammdaten eines Patienten. Diese können daher als Input zu dem von Sweeney entwickelten Verfahren verwendet werden, um die gängigen anonymisierungsrelevanten Begriffe (Personen, Adressen, Abteilungsnamen etc.) im viel unzugänglicheren unstrukturierten Teil eines medizinischen Dokuments zu suchen und zu erkennen.
Viel schwieriger ist allerdings die Identifizierung von Begriffen und Daten, die sich nicht aus den strukturierten Stammdaten oder Dateiköpfen ableiten lassen und die im unstrukturierten Freitext eines medizinischen Dokuments (z.B. Arztbrief, OP-Protokoll, pathologischer Befund) durch eine große lexikalische Vielfalt gekennzeichnet sind. Dazu gehören u.a. Datums- und Zeitangaben aller Art (z.B. *„ED 9/05“*, *„im Mai 2004“*, *„vom 2. bis 6.8.03“*, *„vor 3 Wochen“*, etc.), die beispielsweise im Zusammenhang mit Diagnosen durchaus Rückschlüsse auf die Identität eines Patienten zulassen und die wir deswegen exemplarisch für unsere Studie ausgewählt haben. Wir stellen hier eine Methodik vor, wie mit Hilfe von computerlinguistischen Methoden aus dem Bereich der automatischen Begriffserkennung effektiv anonymisierungsrelevante Datums- und Zeitangaben erkannt und wie die notwendigen Trainingsdaten für das zugrunde liegenden maschinellen Lernverfahren effizient bereitgestellt werden können.
Material und Methoden. Unser Verfahren zur Begriffserkennung beruht auf neuesten Methoden des maschinellen Lernens (Conditional Random Fields – CRF), wie sie in der statistisch orientierten Computerlinguistik angewandt werden [2]. Dabei werden anhand eines Lernverfahrens aus Trainingsdaten Muster abgeleitet (Klassifizierer) anhand derer dann die Wörter eines beliebigen Satzes automatisch mit den (gelernten) Markern versehen werden können. Diese Trainingsdaten, in denen die zu erkennenden Begriffe markiert und semantisch typisiert sind (z.B. <zeitangabe>im Mai 2004</zeitangabe>), müssen allerdings von menschlichen Annotatoren zunächst erstellt werden. Stellt man solchen maschinellen Lernverfahren umfangreiche und qualitativ hochwertige Trainingsdaten zur Verfügung, so zeigen sie sich als äußerst robust (z.B. gegenüber Rechtschreibfehlern und anderen orthographischen Variationen) und weisen außerordentlich hohen Erkennungsrate auf. Das Erstellen solcher textbasierter Trainingsdaten ist allerdings sehr zeit- und arbeitsaufwändig. Daher ist es oft äußerst schwierig, solchen Lernverfahren ausreichend annotiertes Trainingsmaterial zur Verfügung zu stellen, was Auswirkungen auf deren Performanz hat. Erschwerend kommt hinzu, dass selbst große Textmengen oft nur eine geringe Dichte an relevanten Begriffen, die als positive Lernbeispiele dienen können, aufweisen. So müssen Annotatoren zuweilen sehr große Textmengen sichten und annotieren, um eine hinreichend große Anzahl positiver Lernbeispiele zu bekommen. Um diesen Flaschenhals zu umgehen, benutzen wir eine besondere Selektionsstrategie, das sogenannte *Active Learning* (AL) [3], mit Hilfe dessen dem menschlichen Annotator in einem iterativen Verfahren gezielt die für das maschinelle Lernverfahren informativsten Textdaten zur Annotation bereitgestellt werden. In jeder AL-Runde wird ein sogenanntes *Komitee* aus Klassifizierern auf unterschiedlichen Teilbereichen der schon annotierten Texte trainiert. Die so unterschiedlich trainierten Klassifizierer werden dann verwendet, um in den noch nicht annotierten Textdaten die zu erkennenden Begriffe (in unserem Fall Datums- und Zeitangaben aller Art) automatisch zu identifizieren. Auf Satzebene werden dann die von jedem Klassifizierer identifizierten Begriffe miteinander verglichen. Die Sätze, die bezüglich der identifizierten Begriffe die höchste Nichtübereinstimmung aufweisen, werden zur nachfolgenden manuellen Annotation selektiert, da sie besonders informativ für das maschinelle Lernen sind. So wird verhindert, dass unnötig viele unin-formative Sätze annotiert werden müssen. Der AL-Prozess wird beendet, sobald keine bzw. nur noch eine sehr geringe Nichtübereinstimmung zwischen den Klassifizierern besteht.
In unserem Experiment haben wir eine heterogene Textmenge klinischer Dokumenten (Arztbriefe, OP-Berichte, Pathologie- und Histologiebefunde), bestehend aus 3.486 Sätzen und insgesamt 50.655 Wörtern, aus dem klinischen FRAMED-Korpus [4] verwendet. In diesen Texten wurden von einem Medizinstudenten manuell alle vorkommenden Datums- und Zeitangaben nach vorgegebenen Richtlinien annotiert. Das so annotierte Korpus wurde im Verhältnis 70:30 in ein AL-Simulationskorpus (2.440 Sätze) und einen Goldstandard (1.046 Sätze) aufgeteilt. In jeder AL-Runde wurde ein Komitee aus 3 Klassifizierern auf dem schon annotierten Teil des Simulationskorpus trainiert. Die zehn Sätze mit der höchsten Nichtübereinstimmung wurden zur weiteren simulierten manuellen Annotation bereitgestellt. Anschließend wurde ein weiterer Klassifizierer auf den bis dahin annotierten Daten trainiert und dessen Performanz bezüglich der Identifizierung der Zeit- und Datumsangaben auf dem Goldstandard ermittelt. Die durchschnittliche Nichtübereinstimmung der selektierten Sätze wurde nach jeder Iteration berechnet, da diese ein Terminierungskriterium für das AL darstellt. Dieser Simulationsvorgang wurde fünfmal auf verschiedenen 70:30-Korpussplits wiederholt und die Performanz gemittelt. Neben der AL-Selektion wurde auch eine Zufallsselektion durchgeführt: In jeder Runde wurden 10 Sätze zur weiteren simulierten Annotation zufällig ausgewählt, die Performanz wurde wie bei der AL-Selektion ermittelt.

Abb. 1: Performanz bei AL–Selektion und Zufallsselektion



Ergebnisse. Abb. 1 zeigt, dass bei der AL-Selektion schon nach 700 Sätzen (also 70 AL-Runden) ein Performanzwert von 83,1% F-Maß erreicht wird. Bei einer Zufallsselektion wird dieser Wert erst nach 1.900 Sätzen erreicht. In Abb. 2 sieht man, dass die durchschnittliche Nichtübereinstimmung der selektierten Sätze ab der 70. Runde (also ab 700 Sätzen) gegen null tendiert bzw. auf sehr niedrigem Niveau schwankt, was somit auch ein praktisches Abbruchkriterium für den AL-Prozess darstellt. Bei Verwendung von AL muss daher nur ein Drittel der Sätze im Vergleich zur Zufallselektion annotiert werden, um dieselbe Performanz zu erhalten. Einen nur geringen Performanzgewinn von 2 Prozentpunkten auf 85% F-Maß erhält man dagegen bei Annotation des gesamten Korpus. Somit kann durch Verwendung der AL-Selektionsstrategie der Annotationsaufwand um ca. 70 % reduziert werden.
Diskussion. Obwohl unsere Studie auf einem relativ kleinen Datensatz durchgeführt wurde, konnten wir zeigen, dass mit der AL-Selektionsstrategie sehr effizient annotierte Trainingsdaten für ein maschinelles Lernverfahren zur automatischen



medizinischen Begriffserkennung bereitgestellt werden können. Wir haben unseren Ansatz auf schwer zu erfassenden, weil heterogenen, Datums- und Zeitangaben in klinischen Freitexten getestet und konnten zeigen, dass nur ca. 30% der vorhandenen Textmenge annotieren werden muss, um eine fast äquivalente Performanz zu erreichen wie bei Annotation der kompletten Textdaten. In der Praxis könnte ein solches AL-System innerhalb eines Klinikinformationssystems (KIS) eingesetzt werden. Da der Pool an vorhandenen Texten in einem KIS viel größer und repräsentativer ist, ist davon auszugehen, dass bei Rückgriff auf diese Daten sehr gute Erkennungsraten anonymisierungsrelevanter Begriffe erreicht werden können. Eine innerhalb des KIS durchgeführte AL-basierte manuelle Annotation trägt zum Einen den Anforderungen des Datenschutzes Rechnung, da keine unanonymisierten Daten nach außen gelangen, und stellt zum Anderen eine effiziente Annotationsmethode dar. Den hier vorgestellten Ansatz werden wir auf andere, ebenfalls schwer fassbare medizinische Begriffe wie bspw. Maßangaben, Codes, Diagnosen und Medikamentenamen, erweitern, deren Identifizierung im Text essentiell für die klinische Informationsextraktion (z.B. beim Sammeln von Daten für klinische Studien) ist.

Literatur

[1] Sweeney L. Replacing personally-identifying information in medical records, the Srub system. Proc AMIA Annu Fall Symp, 1996, 333-7.
[2] Lafferty J, McCallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. ICML, 2001, 282-9.
[3] Ngai G, Yarowsky D. Rule Writing or Annotation: Cost-efficient Resource Usage for Base Noun Phrase Chunking. Proceedings of ACL, 2000, 117-125.
[4] Wermter J, Hahn U. Ein annotiertes deutschsprachiges medizinisches Textkorpus. GMDS, 2004, 235-7.