Aus dem Institut für Medizinische Biometrie und Medizinische Informatik
der Albert-Ludwigs-Universität Freiburg im Breisgau

Identifikation anonymisierungsrelevanter Informationen in medizinischen Dokumenten

INAUGURAL-DISSERTATION

zur

Erlangung des Medizinischen Doktorgrades
der Medizinischen Fakultät
der Albert-Ludwigs-Universität Freiburg im Breisgau

Vorgelegt 2008
von Felix Balzer
geboren in Hamburg

Dekan:	Prof. Dr. Christoph Peters
Erstgutachter:	Prof. Dr. Stefan Schulz
Zweitgutachterin:	Prof. Dr. Katharina Nübler-Jung
Jahr der Promotion:	2008

Inhaltsverzeichnis

Abkürzungsverzeichnis

AL	Active Learning
BDSG	Bundesdatenschutzgesetz
CRF	Conditional Random Fields
EXT	automatisch generiertes Datenmaterial
HMM	Hidden Markov Model
IE	Information Extraction
IR	Information Retrieval
LDSG	Landesdatenschutzgesetz
MAN	realistisches, manuell annotiertes Datenmaterial
MUC	Message Understanding Conference
NER	Named Entity Recognition
NLP	Natural Language Processing
PHI	Private Health Information
SVM	Support Vector Machines

1 Einleitung

1.1 Thematik

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. Doch der Datenschutz stellt eine hohe Barriere für die Nutzung dieser Informationen dar. Nur bei einer sichergestellten Anonymisierung dürfen die Daten in anderen Bereichen genutzt werden.

Folgende Szenarien sind in diesem Zusammenhang von Bedeutung:

- Klinische Epidemiologie
  Die statistische Auswertung von medizinischen Dokumenten besitzt einen großen Stellenwert. Nach einer standardisierten Aufbereitung könnten medizinische Daten anonymisiert in Datenbanken einfließen und Wissenschaftlern als Material für retrospektive Studien dienen [Chapman01a]. Dies würde es erlauben, gezielten Fragestellungen (z.B. Wirksamkeit von Behandlungsverfahren, onkologische Statistiken, etc.) nachzugehen [Chapman01b].

- Ausbildung von Medizinstudenten und Weiterbildung von Ärzten
  Die aktuelle Reform der ärztlichen Approbationsordnung sieht verstärkt fallorientiertes Lernen in der Ausbildung der Medizinstudenten vor. Die angehenden Mediziner benötigen dafür Anschauungsmaterial in Form von medizinischen Dokumenten wie Befunddokumentationen, Arztbriefen, etc. Auch im Bereich des E-Learnings, der computergestützten Lehre für Medizinstudenten, besteht ein Bedarf an diesen Daten. Generell bietet E-Learning den Vorteil, dass Studenten orts- und zeitunabhängiger Zugang zu

Lerninhalten geboten werden kann. Auch Ärzte können dieses Verfahren für ihre Fort- und Weiterbildungen nutzen. So sind beispielsweise erste Systeme zur CME[1]-Fortbildung über das Internet bereits gestartet. Da der Einsatz von E-Learning-Methoden in den nächsten Jahren tendenziell zunehmen wird, ist damit zu rechnen, dass Applikationen zur automatisierten Anonymisierung ebenfalls dort Anwendung finden werden [Krüger-Brand02].

- Entwicklung von elektronischen Datenverarbeitungssystemen
  Elektronische Systeme zur Erfassung von Patienten haben einen großen Nutzen [Berner05]. Die riesigen Datenmengen, die sich seit der Einführung von EDV tagtäglich im Krankenhausbetrieb ansammeln, haben dazu Anlass gegeben, mit Methoden der Computerlinguistik nach semantischen Gesichtspunkten Information zu finden und zu extrahieren [Friedman04]. Solche Programme können allerdings nur getestet und weiterentwickelt werden, wenn ausreichend Datenmaterial zum Training zur Verfügung steht. Auch Krankenhausinformationssysteme sind für ihren Betrieb auf personenbezogene Daten angewiesen. Bei Schulungen oder Demonstrationen solcher Programme muss gewährleistet sein, dass keine sensiblen Daten der Öffentlichkeit bekannt werden.

Wenn die Gefahr besteht, dass Dritte Einblick in sensible Daten erhalten können, kann grundsätzlich durch Anonymisierung oder Pseudonymisierung der Datenschutz gewährleistet werden.[2] Alle Textbestandteile, aus denen sich die Identität des Patienten ergibt (z.B. Name, Adresse, etc.), werden dabei entfernt; die nicht-sensiblen Bestandteile eines Dokumentes sind hingegen unverändert. Bei einer Anonymisierung könnte man einen Namen so verändern, dass jeder Grossbuchstabe beispielsweise durch ein „X“ und jeder Kleinbuchstabe durch

[1] Continuing Medical Education bezeichnet die kontinuierliche Fortbildung von Ärzten [Deutsches Ärzteblatt08].

[2] Auf die Begriffe Anonymisierung und Pseudonymisierung im juristischen Sinne wird in Kapitel 1.4 Datenschutz eingegangen.

ein „x“ ersetzt wird, z.B. „Herr M. Meier“ → „Herr X. Xxxxx“. Somit blieben die Merkmale Patientengeschlecht, Länge und Struktur des Namens (z.B. Doppelname, Adelstitel) erhalten. Bei der Pseudonymisierung findet hingegen ein Ersatz durch einen fiktiven Namen oder eine Ziffern-/Buchstabenkombination statt, z.B. wird „Meier“ durch „Schmidt“ ersetzt. In einer geheimen Tabelle kann festgehalten werden, welches Pseudonym welchem Patientennamen entspricht. Es bleibt also ein Bezug zum einzelnen Individuum bestehen, dessen Identität allerdings nicht erkennbar ist [Pommerening95].

Wenn diese Methoden nicht zuverlässig eingesetzt werden können, kann in vielen Fällen Verschlüsselung für Datenschutz sorgen [Dierks04]. Verschlüsselung soll dazu dienen, dass sensible Informationen auch dann unter Geheimhaltung ausgetauscht werden können, wenn Dritte Zugang zum Datenfluss haben [Bellare05]. Dabei wird in der Regel ein Text mit Hilfe eines Verschlüsselungsverfahrens so umgewandelt, dass er ohne Kenntnis des Schlüssels nicht mehr lesbar ist. Eine Verwendung im Sinne der oben geschilderten Szenarien ist also nicht möglich. Auch wenn Verschlüsselungsalgorithmen nach dem aktuellen Stand der Wissenschaft als sicher gelten, kann sich dies im Laufe der Jahre durch höhere Rechnerkapazitäten ändern.[3] In der Medizin wird der Begriff der Verschlüsselung häufig ganz anders gebraucht. Gemeint ist dabei die Angabe von Diagnosen in Form einer Buchstaben-/Zahlenkombination zu epidemiologische Zwecken und Verwaltungsangelegenheiten (z.B. ICD-10-Code K35.0 für eine akute Appendizitis mit diffuser Peritonitis) [World Health Organization90].

Bei dem Begriff der anonymisierungsrelevanten medizinischen Daten handelt es sich um eine äußerst heterogene Gruppe. Diese Vielfalt zeigt sich beispiels-

[3] Die Exhaustionsmethode (engl. brute force attack) ist beispielsweise eine potentielle Gefahr für Verschlüsselungssysteme [Rechenberg06]. Dabei werden alle möglichen Kombinationen ausprobiert, um Daten zu entschlüsseln. Je höher die Rechnerkapazität, desto schneller kann der verwendete Schlüssel gefunden werden.

weise in Arztbriefen, in OP-Protokollen oder Befundberichten. Unter dem Begriff Arztbrief wird im Allgemeinen ein ärztlicher Entlassungsbericht verstanden. Er soll eine abschließende kritische, wissenschaftliche Beurteilung eines Krankheitsbildes und Krankheitsverlaufes enthalten und damit beim Empfänger des Briefes der raschen Orientierung dienen, z.B. bei Wiederaufnahme des Patienten oder bei der Beantwortung auftretender Fragen [Heckl90].

1.2 Problemstellung

Selbst relativ standarisierte Dokumente wie Arztbriefe bergen eine Vielzahl von Tücken, die eine automatisierte Bearbeitung nicht ohne weiteres zulassen. Noch schwieriger stellt sich die Anonymisierung von Dateien da, die weniger strukturiert sind (z.B. Befundberichte aus der Pathologie). Eine Vielzahl anonymisierungsrelevanter Informationen befindet sich typischerweise in den strukturierten Abschnitten medizinischer Dokumente, wie z.B. dem Briefkopf, dem Empfängeradressfeld, etc., und sollte somit leicht zu identifizieren sein. 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 durch eine große sprachliche Vielfalt gekennzeichnet sind. Dazu gehören neben Namen von Patienten, Ärzten, Kliniken, Ortsangaben, Berufsbezeichnungen auch 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.

Ziel der vorliegenden Arbeit soll nicht sein, ein betriebsfähiges Anonymisierungssystem zu schaffen; dafür sei auf [Uzuner07] verwiesen. Vielmehr soll der Fragestellung nachgegangen werden, auf welche Weise effizient Trainingsmaterial für eine zuverlässige Erkennung von Eigennamen in Arztbriefen bereitgestellt werden kann. Die einzelnen Studienziele werden in den Einführungen zu den Experimenten erläutert (Kapitel 3.1 und 4.1).

1.3 Einführung in die automatische Informationsextraktion

Die automatische Informationsextraktion (engl. *information extraction*, IE) bezeichnet die Strukturierung und Kombinierung von Daten aus normalsprachlichen Texten [Dale00]. Normalsprachlich bedeutet in diesem Zusammenhang, dass Texte in alltäglicher, menschlicher Sprache vorliegen und keine Steuerzeichen oder ähnliches enthalten, um sie für Computer verständlich zu machen. Diese Extraktion von Informationen aus Dateien ist ein Teilgebiet der maschinellen Sprachverarbeitung (engl. *natural language processing*, NLP), einem Zweig der Wissenschaften Künstliche Intelligenz und Computerlinguistik, der sich mit der automatischen Verarbeitung von menschlicher Sprache beschäftigt. Es bezeichnet computergestützte Verfahren, die Phänomene der gesprochenen oder geschriebenen menschlichen Sprache erkennen und analysieren [Manning05].

Ein IE-System verwendet als Quelle einen Text in menschlicher Sprache. Als Resultat wird daraus ein eindeutiges, nach vorgegebenen Richtlinien strukturiertes Dokument erzeugt. Dieses kann dann direkt dem Anwender am Monitor angezeigt werden oder zur weiteren Verarbeitung in einer Datenbank gespeichert werden [Cunningham97].

In diesem Zusammenhang bietet sich ein Vergleich von *information extraction* und *information retrieval (IR)* Systemen an [ebenda]. Bei einem IR-System, z.B. einer Suchmaschine im Internet, gibt man ein oder mehrere Suchbegriffe an, nach denen in der Folge gesucht wird. Ein Anwender würde dann die gefundenen Seiten betrachten und durch das Lesen dieser Seiten an die für ihn relevanten Details gelangen. Anders ist es bei einem IE-System. Der Prozess des Sichtens und der Extraktion der Informationen erfolgt in diesem Fall durch das System. Auf diese Weise können beispielsweise automatisch Tabellen oder Datenbanken mit dem gefundenen Wissen gefüllt werden.

Die ursprünglich 1987 vom US-amerikanischen Verteidigungsministerium gegründete *Message Understanding Conference* (MUC) beschäftigte sich erstmals mit der Evaluation von Methoden zur Informationsextraktion aus Texten [Grishman96]. Der Fokus lag damals auf der Auswertung militärischer Daten. 1991 fand in den Arbeiten der MUC ein Wechsel des verwendeten Textmaterials von Militärberichten zu Zeitungsartikel statt. Folglich fanden die meisten Untersuchungen auf dem Wortschatz allgemeiner Zeitungsartikel statt. Die Analyse fachspezifischer Texte fand es später Beachtung.[4]

Entsprechend der Klassifikation der MUC stellt Namenserkennung (engl. *named entity recognition*, NER) eine Anwendung von IE dar. Sie dient dem Finden und Klassifizieren von Textbestandteilen in bestimmte Kategorien, wie z.B. Namen, Organisationen, Orte, etc.

Techniken der Namenserkennung in biomedizinischen Anwendungen[5] werden mittlerweile interdisziplinär diskutiert [Baud02]. Friedman *et al.* beschreiben die Möglichkeit Informationen aus Patientenakten mithilfe von NLP zu repräsentieren [Friedman99]. Sie haben ebenfalls ein Verfahren entwickelt, ein klinisches Dokument vollständig in Form von Codes abzubilden [Friedman04]. De Bruijn *et al.* [de Bruijn02] geben einen Überblick über derzeit verwendete Techniken in der Medizin.

Zur Strukturierung von Texten sind einzelne Textbestandteile zu klassifizieren [Carstensen04]. Als erster Schritt ist dafür die Unterteilung des Textes in sinnvolle Einheiten notwendig. Diese Segmentierung eines Textes in bestimmte Einheiten mithilfe eines Tokenizers wird als Tokenisierung bezeichnet. Bei den Tokens handelt es sich im Allgemeinen um Wörter. Die Einteilung der Tokens

[4] Die Informationsextraktion aus biomedizinischen Texten war Gegenstand des BioCreAtIvE - Wettbewerbs [Hirschman05]. Hier lag der Fokus auf der Erkennung von Gen- und Proteinnamen; die Extraktion personenbezogener Informationen wurde nicht untersucht.

[5] Hierzu würde auch ein automatischer Anonymisierer zählen.

auf Wortebene nach bestimmten Kriterien, wie Wortarten oder Eigennamen, erfolgt später nach festgelegten Regeln mithilfe von Tags oder Labels[6] durch so genannte Tagger. Die manuelle Zuordnung von Tokens zu bestimmten Tags wird als Annotation bezeichnet. Unter Tagging versteht man einen automatisierten Prozess, der diese Aufgabe erfüllt (siehe Tabelle 1).[7]

Frau Fröhlich wurde am 30.07.2007 stationär aufgenommen.	
Token	(Semantisches) Tag
Frau	patient
Fröhlich	patient
wurde	O
am	O
30.07.2007	date
stationär	O
aufgenommen	O
.	O

Tabelle 1: Beispiel zur Arbeitsweise eines Taggers. Im vorliegenden Fall sollen Eigennamen (engl. *named entities*) und Zeitangaben erkannt und klassifiziert werden. Andere Anwendungsbeispiele wären z.B. die Erkennung von Wortarten oder numerischen Angaben. Nach Unterteilung des dargestellten Satzes in seine Bestandteile auf Wortebene durch einen Tokenizer werden die einzelnen Tokens mit einem Tag versehen (Tagging). In diesem Beispiel sollen Patientennamen und Datumsangaben gefunden und als solche gekennzeichnet werden. Wird ein Token als nicht relevant eingestuft, erfolgt als Ausgabe „O“.

[6] *tag* und *label*, englisch für Kennzeichnung

[7] Eine detaillierte Betrachtung der Klassifizierung erfolgt in Kapitel 2.2.1.

Ist ein Token einmal als relevant für die Anonymisierung markiert, ist es programmiertechnisch relativ einfach, es zu extrahieren bzw. zu eliminieren. Für den Schritt der Erkennung eines Tokens als relevanten Bestandteil gibt es mehrere Ansätze. Regelbasierte Verfahren gehen nach bestimmten Kriterien vor, um ein Token als bedeutsam zu identifizieren (siehe Tabelle 2).

Der diensthabende Arzt Dr. Quasem ordnete die Durchführung eines CT an.	
Markiere Token als Arztnamen, wenn	– es einer der Zeichenketten [Doktor, Dr., Dr. med., Professor, Prof., Prof. Dr., Prof. Dr. med.] folgt – es mit einem Großbuchstaben beginnt – ungleich der Zeichenketten [med., hc., habil.] ist

Tabelle 2: Beispiel für einen regelbasierten Tagger. In dem aufgeführten Beispiel wird der im Satz enthaltene Arztname korrekt erkannt. Sollte aber beispielsweise ein Patient einen Doktortitel führen und wird dieser so im Arztbrief gefunden, würde der Tagger versagen.

Alternativ zu regelbasierten Taggern gibt es statistische Verfahren. Diese berechnen die Wahrscheinlichkeit, mit der eine Zeichenfolge einer bestimmten Kategorie zuzuordnen ist. Dafür ist eine Vorbereitung erforderlich, in der mithilfe von Trainingsmaterial dem Tagger „beigebracht“ wird, wie relevante Informationen in Texten gefunden werden können. Das Erstellen solcher text-basierter Trainingsdaten ist allerdings sehr zeit- und arbeitsaufwändig, zumal eine manuelle Anonymisierung bereits erfolgt sein muss. Erschwerend kommt hinzu, dass selbst große Textmengen oft nur eine geringe Dichte an relevanten Begriffen, die als positive Lernbeispiele dienen können, aufweisen. So müssen menschliche Annotatoren zuweilen sehr große Textmengen sichten und

annotieren, um eine hinreichend große Anzahl positiver Lernbeispiele zu bekommen.

1.4 Datenschutz

Die gesetzlichen Regelungen zum Datenschutz sind in Deutschland im Wesentlichen im Bundesdatenschutzgesetz (BDSG) und in den Landesdatenschutzgesetzen (LDSG) enthalten, wobei das Prinzip der Subsidiarität gilt. Wann immer es eine konkrete Rechtsvorschrift gibt, tritt das allgemeine Datenschutzrecht in den Hintergrund. Das BDSG regelt die Behandlung sensibler Daten in bundesunmittelbaren Einrichtungen (öffentlicher Bereich: z.B. Bundeswehrkrankenhaus), privatrechtlich organisierten Krankenhäusern und bundesweit tätigen Krankenkassen. Die Landesdatenschutzgesetze gelten für öffentliche Krankenhäuser, die Landesversicherungsanstalt und z.B. die AOK in dem jeweiligen Bundesland [Bake04].

Der Arzt muss dafür Sorge tragen, dass die ihm anvertrauten Informationen nicht in die Hände Unbefugter gelangen. Zunächst einmal ist jede Erhebung, Verarbeitung und Nutzung personenbezogener Daten verboten. Diese sind nach § 3 Absatz 1 BDSG folgendermaßen definiert:

> Personenbezogene Daten sind Einzelangaben über persönliche oder sachliche Verhältnisse einer bestimmten oder bestimmbaren natürlichen Person (Betroffener).

Erst aufgrund eines Gesetzes, eines Vertrages oder eines vertragsähnlichen Vertrauensverhältnisses mit dem Betroffenen (z.B. Arzt-Patient-Beziehung) wird das Arbeiten mit den Daten gerechtfertigt [ebenda].

Im auf die Medizin angewandten Datenschutz gelten die allgemeinen Grundsätze, dass nur solche Daten erhoben werden dürfen, die zur Durchführung der

Behandlung erforderlich sind (Grundsatz der Erforderlichkeit), und dass der Umfang der verarbeiteten Daten so gering wie möglich gehalten werden soll (Grundsatz der Datenvermeidung bzw. Datensparsamkeit). In diesem Zusammenhang ist es vom Gesetzgeber vorgesehen, Daten zu anonymisieren bzw. zu pseudonymisieren, falls dies möglich ist.[8] Ferner werden solche Daten unterschieden, die der Verwaltung und Abrechnung dienen und solche, deren medizinischer Inhalt der Behandlung dient [ebenda].

Für die nähere Betrachtung ist Präzisierung und Abgrenzung der Begrifflichkeiten unumgänglich. Im § 3 des BDSG werden die wichtigsten Begriffe definiert. Im Vordergrund stehen hierbei die personenbezogenen Daten, den „Einzelangaben über persönliche oder sachliche Verhältnisse einer bestimmten oder bestimmbaren natürlichen Person“ (§ 3, Abs. 1). Die natürliche Person wird vom Gesetz dabei als „Betroffener“ definiert. Zu den persönlichen Verhältnissen zählen beispielsweise der Name, das Geburtsdatum, der Familienstand, etc. Angaben zu Eigentumsverhältnissen und ähnlichem zählen zu den sachlichen Verhältnissen. Diese Trennung ist im vorliegenden Fall aber auch ohne Belang, da persönliche wie auch sachliche Verhältnisse gleichermaßen durch das BDSG geschützt sind.

In Absatz 9 des Gesetzes wird ein Grundkatalog besonders sensitiver Daten festgelegt. Hierzu gehören Daten über die ethnische Herkunft, politische Meinung, religiöse oder philosophische Überzeugung, Gewerkschaftszugehörigkeit, Gesundheit oder Sexualleben. Der Punkt Gesundheit schließt hier Informationen über die frühere, derzeitige und zukünftige physische und geistige Gesundheit sowie Daten über Drogen- und Alkoholmissbrauch mit ein. Bedenkt man, wie häufig Menschen aufgrund der oben genannten Eigenschaften diskriminiert wurden und werden, erscheint die Hervorhebung eines besonderen Schutzes als nahe liegend.

[8] §§ 3 Abs. 6 und 6a, 3a BDSG

Lässt sich eine Person aus vorliegenden Daten eindeutig identifizieren, dann gilt die Person im juristischen Sinne als *bestimmt*. Das Patientenetikett auf einem EKG-Ausdruck oder auf einem Röntgenbild, welches Name und Geburtsdatum beinhaltet, wäre dafür ein Beispiel. Aus Sicht der Informatik befindet hier sich der Schlüssel unmittelbar im Datensatz. Bestimm*bar* ist hingegen eine Person, wenn Zusatzwissen erforderlich ist, um aus vorliegenden Daten eine Person eindeutig zu identifizieren. Die Frage, ob eine Person bestimmbar ist, ist natürlich relativ. Personenbezug ist also nur in Relation zu konkretem personenbezogenem Zusatzwissen möglich. [Tinnefeld05]

Nach § 3 Absatz 6 des Bundesdatenschutzgesetz gelten personenbezogene Daten dann als anonymisiert, wenn sie derart verändert worden sind, dass [sie] nicht mehr oder nur noch mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft einer […] Person zugeordnet werden können.[9] Dieses kann durch das Weglassen bestimmter Daten (z.B. Name, Wohnort, etc.) erreicht werden. Im selben Paragraph wird im Absatz 6a auch der Begriff des Pseudonymisierens definiert. Dabei handelt es sich um das Ersetzen des Namens und anderer Identifikationsmerkmale durch ein Kennzeichen zu dem Zweck, die Bestimmung des Betroffenen auszuschließen oder wesentlich zu erschweren.[10]

Sobald personenbezogene Daten anonymisiert oder pseudonymisiert sind, ein unmittelbarer Personenbezug also nicht mehr herstellbar ist, greifen die Datenschutzregeln des BDSG nicht mehr. Nach Tinnefeld at al. ist das Ziel der Anonymisierung [Tinnefeld05]

[9] Vergleiche auch § 16 Abs. 6 BStatG.

[10] Für den genauen Wortlaut im Gesetzbuch sei auf Anhang 7.4 Bundesdatenschutzgesetz (Auszug) hingewiesen.

- die Beziehbarkeit eines Datums[11] oder eines Datensatzes auf eine bestimmte oder bestimmbare Person zu beseitigen,
- aber den inhaltlichen Aussagehalt für die nicht mehr identifizierbare Person zu erhalten.

Bei jeder Anonymisierung muss eine Abschätzung unternommen werden, ob die Daten re-individualisiert werden können, also ob die Möglichkeit besteht, einen anonymisierten Datensatz doch einer Person eindeutig zuzuordnen. Da dieses Risiko nie absolut ausgeschlossen werden kann, spricht man von faktischer Anonymisierung.

[11] Der Begriff *Datum* ist hier als Informationseinheit, nicht als Zeitangabe zu verstehen.

2 Methoden (allgemein)

2.1 Aufbau

Da das Wesentliche der automatisierten Anonymisierung von Medizintexten die Erkennung anonymisierungsrelevanter Informationen ist, beschäftigt sich die vorliegende Arbeit mit der Frage, auf welche Weise ein System effizient zur korrekten Klassifizierung trainiert werden kann.

Die durchgeführten Experimente gliedern sich in zwei Teile. Beide Teile können als voneinander unabhängige Studien betrachtet werden, deren Methodik jedoch Gemeinsamkeiten aufweist. In der ersten Studie wurde der Frage nachgegangen, wie durch den Einsatz von *active learning* der Aufwand zur Akquirierung von manuell annotiertem Trainingsmaterial auf ein Minimum reduziert werden kann. Die zweite Studie beschäftigt sich mit der Aufgabe, die manuelle Bereitstellung von Trainingsmaterial dadurch zu reduzieren, indem annotiertes Material automatisch generiert wird.

Beide Studien erfolgten in Kooperation mit der Abteilung für Medizinische Informatik der Albert-Ludwigs-Universität Freiburg[12] und dem Jena University Language & Information Engineering Lab[13].

[12] http://www.uniklinik-freiburg.de/medinf/live/abteilung.html

[13] http://www.coling-uni-jena.de

2.2 Arbeitsweise der verwendeten Techniken

2.2.1 Allgemeine Funktionsweise von Klassifizierern zur Textkategorisierung

Klassifizierer erlauben die Einteilung von Texten oder Textteilen in verschiedene Kategorien. Für eine solche programmgesteuerte Analyse von medizinischen Texten wie Arztbriefen sind mehrere Verarbeitungsschritte erforderlich [Tomanek07a].

Ein *sentence splitter* erhält dafür als Eingabe einen kompletten Text, der dann in Sequenzen unterteilt wird. Diese Sequenzen können beispielsweise in ihrer Dimension einem Satz entsprechen. Hierzu wird der Text zunächst in kleine Einheiten unterteilt und zwar immer an den Stellen getrennt, wo sich ein Leerzeichen befindet. Wenn eine solche Einheit ein so genanntes *sentence boundary symbol*[14] enthält, muss der Algorithmus entscheiden, ob es sich tatsächlich um das Ende eines Satzes handelt oder etwa nur eine Abkürzung, wie z.B: „usw.“. In einem späteren Arbeitsschritt dienen diese Textsequenzen einem Tokenizer als Quelle, um die einzelnen Tokens zu identifizieren. Wichtig ist hier, dass nicht nur Wörter, sondern auch Satzzeichen als Tokens betrachtet werden. So wird beispielsweise die Zeichenkette „sagte, dass“ in die Tokens [sagte] [,] [dass] zerlegt.

Zur Beschreibung der Tokens werden die folgenden Attribute[15] herangezogen, welche den verwendeten Tools als Entscheidungshilfe dienen [Tomanek07b]:

- *orthographical*
  Auf dieser Ebene erfolgen dichotome Beschreibungen der Orthographie anhand von regulären Ausdrücken[16].

[14] Dazu zählen beispielsweise Punkte, Fragezeichen, Ausrufezeichen.

[15] in englischsprachiger Literatur als *Feature* bezeichnet

[16] Reguläre Ausdrücke dienen der Beschreibung von Mengen bzw. Untermengen von

New RegexMatches(…)	Regulärer Ausdruck positiv, wenn
"INITCAPS", Pattern.compile("[A-Z].*")	das erste Zeichen ein Großbuchstabe ist
"ALLCAPS", Pattern.compile("[A-Z]+")	alle Zeichen Großbuchstaben sind
"CAPSMIX", Pattern.compile("[A-Za-z]+")	es sich um Groß- oder Kleinbuchstaben handelt
("PUNCTUATION", Pattern.compile("[,.;:?!-+]")	ein Satzzeichen vorliegt
SINGLEDIGIT", Pattern.compile("[0-9]")	das Token aus einer Ziffer besteht
DOUBLEDIGIT", Pattern.compile("[0-9]0-9]")	das Token aus zwei Ziffern besteht

Tabelle 3: Syntax des MALLET-Toolkits zur Beschreibung der Attribut-Klasse *orthographical*

- *lexical and morphological*

In dieser Attribut-Klasse wird beschrieben, ob das betrachtete Token ein Prä- oder Suffix enthält. Außerdem findet eine Transformation zur Grundform des Wortes statt, z.B. Singular bei Nomen.

- *syntactic*

Hierbei wird das Token nach syntaktischen Gesichtspunkten untersucht (z.B. Wortart, Stellung im Satz).

- *contextual*

Die Kenntnis, ob es sich bei einem Wort um das Subjekt oder Objekt handelt, bestimmt zu einem großen Teil die Bedeutung des Satzes. Deshalb werden auf dieser Ebene die Attribute der benachbarten Tokens untersucht.

Zeichenketten mit Hilfe bestimmter syntaktischer Regeln. Wie ein Filter kann durch Angabe eines regulären Ausdrucks ein Dokument nach einem Muster durchsucht werden. Ferner erlauben reguläre Ausdrücke das Erzeugen einer Menge von Wörtern nach vorher festgelegten Regeln im Sinne einer Schablone [Rechenberg06].

Für die durchgeführten Experimente wurden die von McCallum unter der Bezeichnung „MALLET: A Machine Learning for Language Toolkit. 2002“[17] entwickelten Programme genutzt. Die einzelnen Tools sind allesamt in Java geschrieben und stehen als Open Source Software[18] zu Verfügung.

2.2.2 Conditional Random Fields (CRF)

Im Bereich des maschinellen Lernens ist eine Vielzahl von Techniken zur Klassifizierung bekannt. Dazu gehören beispielsweise der Bayes-Klassifikator[19] und *Support Vector Machines* (SVM)[20]. Das *Hidden Markov Model* (HMM) stellt eine weitere Methode zur Erkennung von bestimmten Mustern dar. Das zu modellierende System wird dabei als Markov Prozess bezeichnet, wobei anhand der bekannten Parameter die unbekannten Parameter des Systems ermittelt werden sollen [Luger03]. Conditional Random Fields kommen bei ähnlichen Fragestellungen zum Einsatz wie HMM, wobei Letztere in ihrem konzeptionellen Aufbau wesentlich einfacher sind. Ein CRF kann als ungerichteter Graph verstanden werden, der zur Bestimmung von Klassenzugehörigkeiten aufgrund gemachter Beobachtungen benutzt wird.[21] Diese häufig in der statistisch orientierten Computerlinguistik angewandte Methode zeichnet sich durch außerordentlich hohe Erkennungsraten[22] sowie Robustheit

[17] http://mallet.cs.umass.edu

[18] http://www.opensource.org/licenses/cpl.php

[19] Der Bayes-Klassifikator ist eine mathematische Funktion, die jedes Objekt einer Klasse zuordnet, zu der es mit der größten Wahrscheinlichkeit gehört, oder bei der durch die Einordnung die wenigsten Kosten entstehen. Grundlage ist das nach dem englischen Mathematiker Thomas Bayes benannte Bayestheorem [Rechenberg06].

[20] Bei diesem mathematischen Verfahren zur Mustererkennung wird eine Menge von Objekten so in Klassen eingeteilt, dass um die Klassengrenzen herum ein möglichst breiter Bereich frei von Objekten bleibt [Cortes04].

[21] Für die theoretischen Grundlagen sei auf [Wallach04] verwiesen.

[22] Mit einer Fehlerquote von 5,55% bei der Bestimmung von Wortarten eines 1,1 Millionen Wörter großen Korpus, wobei 50% des Korpus als Trainingsmaterial genutzt wurde, erwiesen sich CRF als vorteilhafter im Vergleich zu Hidden Markov Models und Maximum Entropy

gegenüber Rechtschreibfehlern und anderen orthographischen Variationen aus [Lafferty01]. Diverse Probleme von HMM können durch den Einsatz von CRF umgegangen werden [Wallach04].

Zur automatischen Annotation von Wörtern eines beliebigen Satzes sind so genannte Muster erforderlich. Diese müssen zuvor anhand eines Lernverfahrens aus Trainingsdaten abgeleitet werden. Die Trainingsdaten enthalten von einem menschlichen Annotator hinterlegte Informationen zu der semantischen Typisierung der einzelnen Tokens (z.B. <zeitangabe>im Mai 2004</zeitangabe>).

2.2.3 Active Learning (AL)

Die manuelle Annotation von Texten für das Training von Algorithmen des maschinellen Lernens erfordert hohen personellen Aufwand. Eine Möglichkeit der Optimierung ist Active Learning. Ngai *et al.* haben ein Modell vorgestellt, welches effektiv im Bereich des NLP genutzt werden kann [Ngai00]. Durch halbautomatische Annotation ist es möglich, eine vergleichbare Performanz mit weniger Annotationsaufwand zu erreichen.

Die Ausgangslage ist, dass relativ wenig annotiertes Material und dementsprechend eine große Menge an nicht annotiertem Material vorliegt. Nachdem ein Tagger mit dem vorhandenen, bereits annotierten Material trainiert worden ist und sozusagen ein Grundstock von Wissen akquiriert wurde, ist die Idee, in einer zweiten Phase vom System ausgewählte, schwierige Sätze eines nicht annotierten Korpus[23] einer realen Person zur Annotation vorzulegen. Bei der Auswahl von Sätzen durch den Computer liegt der Gedanke zugrunde, dass je ungewisser eine ausgewählte Textstelle und je unklarer der Kontext ist, es desto sinnvoller wäre, gerade diese Textstelle von einem Menschen korrekt

Markov Models [Lafferty01].

[23] Textkorpus, das aus einer Sammlung von Texten besteht

annotieren zu lassen. Auf diese Weise müssen also nur diejenigen Korpusanteile annotiert werden, die für den Tagger einen Informationsgewinn darstellen. Der Aufwand, der für die manuelle Annotation von biomedizinischen Texten aufgewendet wurden, konnten durch dieses Modell um 86% reduziert werden.

Bei den drei grundlegenden Kriterien bei der Bewertung von Textstellen handelt es sich in dem Modell um *informativeness*, *representativeness* und *diversity*. Ersteres bezieht sich darauf, ob die Textstelle im vorliegenden Kontext eine Neuerung darstellt, das heißt als nicht gewiss klassifiziert werden kann. Beim Kriterium der *representativeness* geht es darum, aus einer Reihe von Beispielen gerade dasjenige auswählen, welches am ähnlichsten zu denen ist, die im Dokument vorkommen. Schließlich wird durch das *diversity* Kriterium bei der Auswahl der zu annotierenden Textbestandteilen für Vielfalt gesorgt, sodass nicht zu oft sich wiederholende Beispiele aus dem Korpus ausgewählt werden [Dan04].

2.3 Zielparameter der Studien

In den hier beschriebenen Studien wird von verschiedenen Systemen Datenmaterial generiert, welches zu Zwecken der Evaluation mit zuvor festgelegten Goldstandards verglichen wird, so dass Aussagen über die Güte des generierten Materials getroffen werden können. Hierbei werden jeweils Dateipaare betrachtet, die alle Tokens zusammen mit dem jeweiligen Tag eines bestimmten Korpus enthalten. Die eine Datei (Goldstandard) stellt das Musterbeispiel da, wobei man davon ausgehen kann, dass alle Tokens weitestgehend mit den korrekten Tags versehen sind (meist durch manuelle Annotierung). Die andere Datei (Evaluationskorpus) enthält das Ergebnis eines Klassifikationsversuchs durch einen Tagger.

Der Vergleich zwischen Goldstandard und automatischem Verfahren erfolgt im Allgemeinen anhand der im Folgenden beschriebenen Parameter, die besonders im *information retrieval* Anwendung finden.

2.3.1 Precision

Die Precision bezeichnet die Wahrscheinlichkeit, mit der ein gefundenes Objekt relevant ist [Rijbergen79]. Dabei kann es sich beispielsweise um ein Dokument, einen Datensatz oder, wie in diesem Fall, um ein Token handeln.

$$\text{Precision} = \frac{|\text{relevante Objekte} \cap \text{gefundene Objekte}|}{|\text{gefundene Objekte}|}$$

Bezogen auf den in dieser Arbeit durchgeführten Vergleich eines Goldstandards und eines automatisch generierten Evaluationskorpus, stellen die relevanten Dokumente die im Goldstandard enthaltenen anonymisierungsrelevanten Information dar, hier als $E_{Goldstandard}$ bezeichnet. $E_{Evaluation}$ enthält die im Evaluationskorpus tatsächlich gefundenen anonymisierungsrelevanten Informationen. Die Schnittmenge beider Mengen enthält die Elemente, die nach dem Goldstandard im Evaluationskorpus als korrekt identifiziert wurden. Der Quotient aus dem Betrag dieser Schnittmenge und der Zahl der im Evaluationskorpus enthaltenen Elemente $E_{Evaluation}$ gibt die Precision wieder. Diese Wahrscheinlichkeit, mit der anonymisierungsrelevante Informationen gefunden worden sind, wird auf die Anzahl aller gefundenen Informationen im Evaluationskorpus bezogen. Bei dieser Bezugsgröße ist es nicht von belang, ob die jedoch um korrekt oder inkorrekt vom Tagger klassifiziert wurden, ist unerheblich. Wenn eine berechnete Precision einen relativ niedrigen Wert hat, ist dies ein Zeichen dafür, dass zu viele Informationen im Evaluationskorpus fälschlicherweise als relevant markiert wurden.

$$\text{Precision} = \frac{\left|E_{\text{Goldstandard}} \cap E_{\text{Evaluation}}\right|}{\left|E_{\text{Evaluation}}\right|}$$

2.3.2 Recall

Der Recall bezeichnet die Wahrscheinlichkeit, mit der ein relevantes Dokument gefunden wird [Rijbergen79].

$$\text{Recall} = \frac{\left|\text{relevante Objekte} \cap \text{gefundene Objekte}\right|}{\text{relevante Objekte}}$$

Dementsprechend stellt die Schnittmenge der im Goldstandard enthaltenen Elemente $E_{\text{Goldstandard}}$ und der im Evaluationskorpus vorkommenden Elementen $E_{\text{Evaluation}}$ die Menge der Elemente dar, die im Evaluationskorpus nach dem Goldstandard als korrekt identifiziert wurden. Die Berechnung des Quotienten aus dem Betrag dieser Schnittmenge und der Zahl der im Goldstandard enthaltenen Elemente $E_{\text{Goldstandard}}$ ergibt den Recall. Diese Wahrscheinlichkeit gibt darüber Auskunft, wie viele der anonymisierungsrelevanten Informationen tatsächlich gefunden wurden. Ein niedriger Wert lässt vermuten, dass zu viele Tokens im Evaluationskorpus fälschlicherweise als nicht relevant markiert wurden.

$$\text{Recall} = \frac{\left|E_{\text{Goldstandard}} \cap E_{\text{Evaluation}}\right|}{\left|E_{\text{Goldstandard}}\right|}$$

2.3.3 F-Measure

Das F-Measure oder F-Maß, auch als *balanced F-score* bezeichnet, stellt das harmonische Mittel von Precision und Recall dar [Rijbergen79]. Das harmonische Mittel ist in der Mathematik definiert als

$$\mathrm{H} = \frac{2a_1a_2}{a_1 + a_2}$$

Setzt man in diese Gleichung die Variablen für Precision und Recall ein, erhält man

$$\mathrm{F} = \frac{2\,(\mathrm{precision} * \mathrm{recall})}{(\mathrm{precision} + \mathrm{recall})}$$

Es erlaubt mit einem Maß gleichzeitig Auskunft darüber zugeben, mit welcher Wahrscheinlichkeit relevante Informationen gefunden wurden unter Einbeziehung der fehlerhaften Erkennungen in Form von falsch positiven Klassifizierungen (niedrige Precision bei fälschlicherweise relevant markierten Textbestandteilen) und falsch negativen Klassifizierungen (niedriger Recall bei fälschlicherweise nicht relevant markierten Textbestandteilen).

3 Experiment I: Erkennung von Datums- und Zeitangaben

3.1 Einführung

Da Datums- und Zeitangaben wesentlich homogener in ihrer Art als Namen und andere persönlichen Daten sind, hat es sich angeboten, im Vorfeld mit diesen Angaben zu experimentieren und eine Strategie für eine effektive Klassifizierung zu entwickeln. Grundidee war, dass mit einem möglichst geringen Einsatz von menschlichen Annotatoren viel Trainingsmaterial für ein Klassifizierungssystem bereitgestellt werden sollte. Dazu lag der Fokus der Studie auf Active Learning, um den manuellen Annotationsaufwand auf ein Minimum zu reduzieren.

Nachdem alle Datums- und Zeitangaben manuell annotiert wurden, fanden die Daten in einer Simulation Verwendung, die durch Anwendung einer Active Learning - Strategie gesteuert wurde. Die Klassifizierung erfolgte anhand von CRF.

3.2 Material und Methoden

3.2.1 Zusammensetzung des Korpus und manuelle Annotation

In diesem Experiment wurde eine heterogene Textmenge klinischer Dokumente (Arztbriefe, OP-Berichte, Pathologie- und Histologiebefunde), bestehend aus 3.486 Sätzen und insgesamt 50.655 Wörtern, aus dem klinischen FRAMED-Korpus [Wermter04] verwendet. In diesen Texten wurden von zwei Medizinstudenten manuell alle vorkommenden Datums- und Zeitangaben nach vorgegebenen Richtlinien[24] annotiert. Dazu wurde die von der EML Research

[24] siehe Anhang 7.3

gGmbH[25] entwickelten Software mmax2 eingesetzt. Das Programm erlaubt die Festlegung einer beliebigen Anzahl von Wörtern als so genannte *markables*, welche dann einer vorher festgelegten *NamedEntity*, wie z.B. *person*, zugewiesen werden können (Abb. 1).

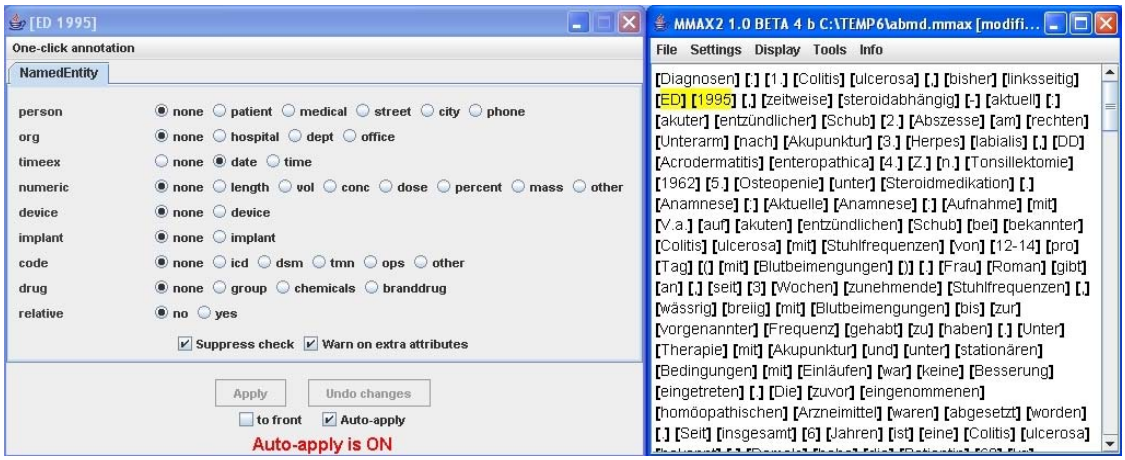

Abbildung 1: Anwendung der Software mmax2 zur manuellen Annotation der klinischen Texte

3.2.2 Einsatz von Conditional Random Fields und Active Learning

Bevor die einzelnen Textbestandteile des Evaluationskorpus anhand von CRF mit Tags versehen werden konnten, musste zuerst ein Trainingskorpus akquiriert werden. Hierzu wurde AL eingesetzt: Ein iteratives[26] Verfahren stellte einer realen Person gezielt die für das maschinelle Lernverfahren informativsten Textdaten zur Annotation bereit.

In jeder AL-Runde wurde ein so genanntes Komitee aus Klassifizierern auf unterschiedlichen Teilbereichen der schon annotierten Texte trainiert. Die so unterschiedlich trainierten Klassifizierer wurden dann verwendet, um in den noch nicht annotierten Textdaten die zu erkennenden Begriffe (in diesem Fall

[25] http://mmax.eml-research.de

[26] wiederholend; in Programmiersprachen wird beispielsweise in einer FOR-Schleife iterativ auf Datenstrukturen zugegriffen

Datums- und Zeitangaben aller Art) automatisch zu identifizieren. Auf Satzebene wurden dann die von jedem Klassifizierer identifizierten Begriffe miteinander verglichen. Die Sätze, die bezüglich der identifizierten Begriffe die höchste Nichtübereinstimmung aufwiesen, wurden zur nachfolgenden manuellen Annotation selektiert, da sie besonders informativ für das maschinelle Lernen sind. So wurde verhindert, dass unnötig viele uninformative Sätze annotiert werden müssen. Der AL-Prozess wurde beendet, sobald keine bzw. nur noch eine sehr geringe Nichtübereinstimmung zwischen den Klassifizierern bestand.

3.2.3 Simulationsvorgang

Für die Evaluation dieser Selektionsstrategie wurde das gesamte annotierte Korpus 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 drei 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 Active Learning 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 zehn Sätze zur weiteren simulierten Annotation zufällig ausgewählt, die Performanz wurde wie bei der Active Learning-Selektion ermittelt.

3.3 Ergebnisse

Abbildung 2 zeigt, dass bei der AL-Selektion schon nach 700 Sätzen (also 70 AL-Runden) ein Performanzwert von 83,1% F-Measure erreicht wird. Bei einer Zufallsselektion wird dieser Wert erst nach 1.900 Sätzen erreicht.

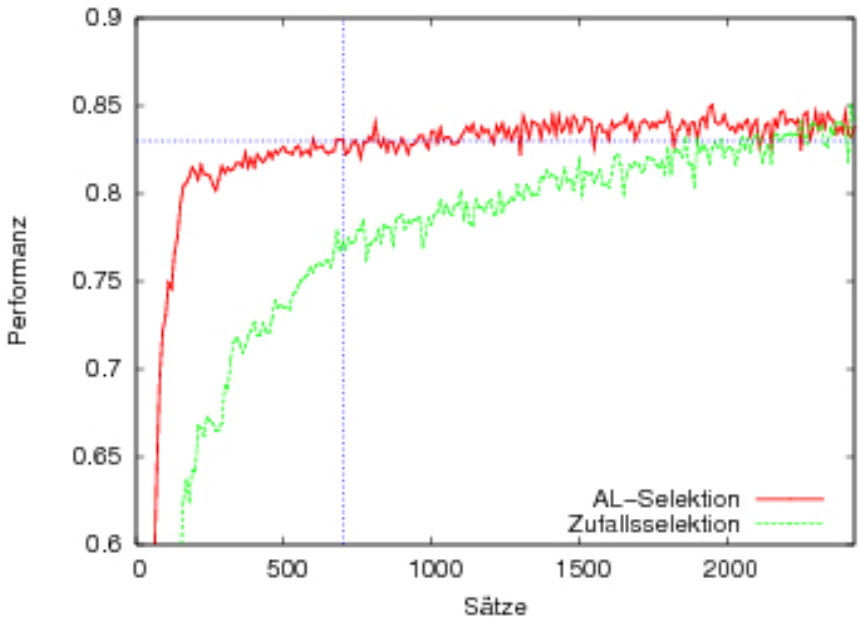

Abbildung 2: Performanz bei AL-Selektion und Zufallsselektion

In Abbildung 3 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 Active Learning muss daher nur ein Drittel der Sätze im Vergleich zur Zufallselektion annotiert werden, um dieselbe Performanz zu erhalten. Einen nur geringen Performanzgewinn von zwei Prozentpunkten auf 85% F-Measure erhält man dagegen bei Annotation des gesamten Korpus. Somit kann durch Verwendung der Active Learning-Selektionsstrategie der Annotationsaufwand um ca. 70% reduziert werden.

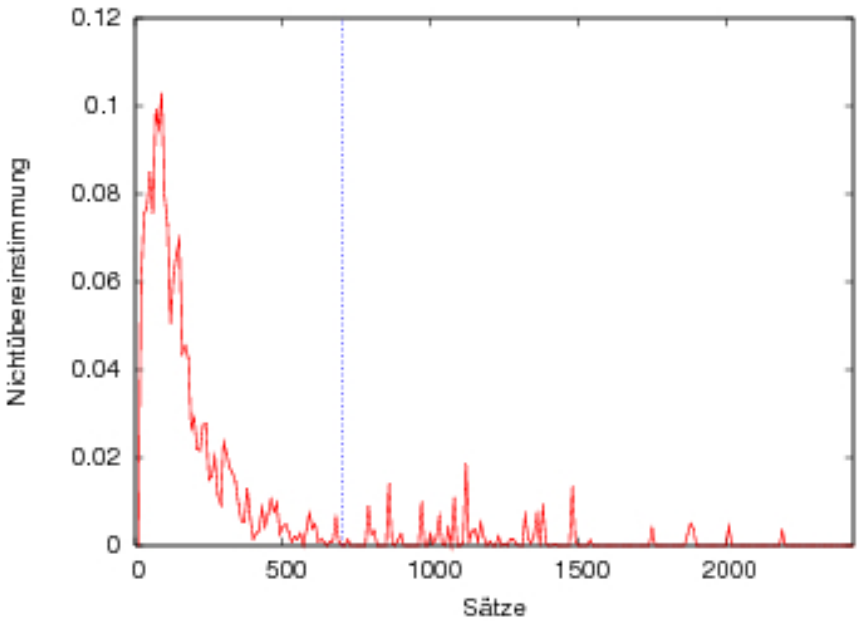

Abbildung 3: Nichtübereinstimmung bei AL-Selektion

4 Experiment II: Erkennung von Eigennamen, Adressen und Institutionen in Arztbriefen

4.1 Einführung

Im vorangegangen Experiment wurde davon ausgegangen, dass Texte mit den zu identifizierenden Begriffen in ausreichender Anzahl vorlagen. Wie eingangs beschrieben ist die Freigabe von Dokumenten für die Forschung, die sensible Informationen enthalten, höchst kompliziert. Die geringe Anzahl der manuell anonymisierten Arztbriefe, die für das Experiment zur Verfügung standen, reichte bei weitem nicht aus, um repräsentative Aussagen über eine effiziente Erkennung der festgelegten anonymisierungsrelevanten Textbestandteile treffen zu können. Deshalb wurde der Fragestellung nachgegangen, ob eine automatische Generierung von fiktiven Arztbriefen möglich ist, um Klassifizierer effektiv zu trainieren. Dazu wurde nicht untersucht, welche Attribute (z.B. Wörterbücher) gesondert zur Erkennung von sensiblen Textbestandteilen herangezogen werden könnten.

Als anonymisierungsrelevant wurden in diesem Experiment folgende Textbestandteile betrachtet. Nähere Erläuterungen zu den Typen finden sich in den Annotationsrichtlinien (siehe Anhang 7.3).

Klasse	**Typ**
person	patient
	medical
address	street
	city
	phone
organisation	hospital
	office
	dept

Tabelle 4: Anonymisierungsrelevante Klassen mit den dazugehörigen Typen

4.2 Material und Methoden

4.2.1 Automatische Generierung von Texten

Zur automatischen Generierung von Arztbriefen diente eine Microsoft Access 2003 - Datenbank (Datenfelder siehe Tabelle 5), die zusammen mit Microsoft Word 2003 eingesetzt wurden. Die Spalten *Vorname*, *Nachname* (ohne Doppelnamen) und *Straße* wurden mit repräsentativen Angaben für eine deutsche Stadt mit ca. einer halben Millionen Einwohner gefüllt. Die Spalte *Titel* enthielt eine Auswahl von Titeln, die häufig im deutschsprachigen Raum anzutreffen sind, wie beispielsweise „Dr." oder „Prof.", aber auch seltenere Titel, wie „Baron" oder „Freiherr". Die Ortsnamen stammten aus einer Zusammenstellung aller deutschen Orte der Firma SeBaWorld[27]. Für die Felder *Anamnese*, *Befund*, *Zusatzuntersuchung*, *Diagnose* und *Epikrise* wurden Textbausteine manuell aus anonymisierten Arztbriefen in die Datenbank eingefügt. Alle Einträge der Datenbank wurden annotiert aufgenommen.

Menge	**Anzahl der Elemente**	**Menge**	**Anzahl der Elemente**
Titel	9	Anamnese	20
Vorname	4462	Befund	17
Nachname	59010	Zusatzuntersuchung	85
Straße	3210	Diagnose	199
Ort	1317	Epikrise	24

Tabelle 5: Vorhandene personenbezogene Textbestandteile in der Datenbank

[27] Deutschland auf einen Blick, http://www.daeb.de

Die Namen wurden nach vorher festgelegten Richtlinien zusammengesetzt. So wurde in fünf Prozent der Fälle ein Doppelname oder ein Name mit dem Zusatz „von“ erstellt. Ebenfalls zu fünf Prozent wurden Titel aus der entsprechenden Tabelle in die generierten Namen hinzugenommen. Die für eine Anschrift noch fehlenden Elemente der Postleitzahl und der Hausnummer wurden anhand von Zufallszahlen erstellt.

In echten Arztbriefen finden sich häufig einfache Referenzen durch den Namen (z.B. „Herr Müller wurde elektiv zur Abklärung einer sonographisch diagnostizierten Raumforderung aufgenommen.“). Dies wurde bei der automatischen Erstellung insofern berücksichtigt, dass Platzhalter in den Textbausteinen der Anamnese, Epikrise etc. durch den generierten Patientennamen ersetzt wurden. Auch das mehrfache Vorkommen von denselben Arztnamen (z.B. Chefarztname im Briefkopf als auch im Unterschriftenblock) wurde bedacht. Im Bestreben, die Arztbriefe so authentisch wie möglich erscheinen zu lassen, wurden Tippfehler in Textbausteinen nicht korrigiert bzw. an bestimmten Stellen nach dem Zufallsprinzip absichtlich erstellt. Ignoriert wurde die Berücksichtigung des Geschlechtes des Patienten bei Referenzen in Textbausteinen. So wurden Bezüge wie „der Patient beklagte…“ im Freitext des Arztbriefes bei einem weiblichen Patientennamen in der Betreffzeile nicht weiter beachtet.

In Microsoft Word wurden 28 Arztbrief-Vorlagen erstellt, die sich alle im Aufbau unterschieden und verschiedene Briefköpfe von fiktiven Kliniken beinhalteten. Für alle in Tabelle 5 vorkommenden Gruppen existierten in den Dateien Platzhalter. Alle variablen Textbestandteile wurden beim Verbinden der Vorlagen mit der Datenbank durch die zufällig ausgewählten Textbausteine ersetzt. Ein Beispiel eines automatisch generierten Arztbriefes ist in Abbildung 4 zu sehen.

<hospital>Universitätsklinikum<city> Brühl</city></hospital>
<street>Koloniestr. 60</street>, <city>97022 Brühl</city>
Pforte: <phone>0256-3978-4732</phone>

<dept>Station XI , Herz- und Gefäßchirurgie</dept>
<medical>Chefarzt: PD Dr. med. K. Garbade</medical>

Sekretariat: <phone>-3845</phone>
Station: <phone>-3842</phone>
Fax: <phone>-8474</phone>
Arzt: <phone>-2947</phone>

Herrn
<medical>Dr. med. M. Bunselmeier</medical>
<street>Ludwigsburger Str. 48</street>
<city>55321 Königsberg</city>

<city>Brühl</city> , <date>28.03.2003</date>

<patient>Stefanski , Helmuth</patient> , geb. <date>19.02.1958</date> in <city> Sierksdorf</city>

#newline-off
Sehr geehrte Frau Kollegin , sehr geehrte Herr Kollege,

folgend erhalten Sie den Bericht über unseren Patienten <patient>Herrn Helmuth Stefanski</patient> , der sich vom <date>14.03.2003</date> bis zum <date>24.03.2003</date> in unserer stationären Betreuung befand .

Diagnosen
1. reaktive Depression

2. Z.n. SM-Implantation <date>03/87</date> , Explantation <date>06/96</date> wegen Elektrodeninfektion

3. CMV-Reaktivierung unter Immunsuppression ICD D-078.5

4. Renale Hypertonie ICD 403.0

5. Zentrales Plattenepithelkarzinom des rechten Lungenoberlappens

Anamnese
Elektive Aufnahme zur Abklärung einer sonographisch diagnostizierten Raumforderung am linken oberen Nierenpol sowie einer langjährigen Leukozytose .

Wohlbefinden .
Bei Selbstmessung nachmittags teilweise erhöhte Blutdruckwerte , morgens Blutdruckwerte im Normbereich .
Seit psychosomatischer Rehabilitation bessere Stimmung , besserer Umgang mit Ängsten .Vorgeschichte :
siehe Diagnosenliste .
Bei Stuhlunregelmässigkeiten , rez. breiigen Stühlen und Blutauflagerungen Koloskopie im <date>Juni d. J.</date> bis auf innere Hämorrhoiden unauffällig .
Seit <date-relative>20 Jahren</date-relative> Leukozytose , Werte um 11 Tsd/µl .
Seit Radiojod-therapie Angstzustände und depressive Stimmung , in ambulanter Psychotherapie .
<date>11/01</date> Diagnose eines seborrhoischen Ekzems nasolabial , Mundwinkel beiderseits und frontal .
Familienanamnese :
Mutter <date-relative>50jährig</date-relative> an Magenkrebs , Vater an Folgen eines Autounfalls verstorben .
Eine Schwester an Brustkrebs erkrankt .
Systemanamnese :
Miktion unauffällig .
Stuhlgang unregelmässig , z.T. flüssig , z.T. geformt .
Gewichtszunahme von 2 kg in den letzten <date-relative>5 Jahren</date-relative> .
Noxen :
20 Zigaretten/Tag , 30 py.
4 Flaschen Bier pro Tag .
Sozialanamnese :
verheiratet , 1 Sohn , Elektroinstalleur .

Befund
<date-relative>71-jähriger</date-relative> Patient in gutem Allgemein- und Ernährungszustand (80 kg , 176 cm) .
RR 120/60 mmHg , Puls 50/min , Temperatur 36,4°C , Schleimhäute feucht , keine Lymphadenopathie tastbar .
Herztöne rein , keine Extratöne , keine vitientypischen Geräusche .
Vesikuläratmen , keine pathologischen Nebengeräusche .
Bauchdecke weich , reizlose Laparotomienarbe im rechten Unterbauch .
Darmgeräusche lebhaft .
Leber unter dem Rippenbogen anstoßend tastbar .
Milz nicht palpabel .

Verlauf und Beurteilung
<patient>Herr Stefanski</patient> wurde aus der <hospital>Stadtklinik <city>Engental</city></hospital> zu uns zur
PTCA einer hochgradigen RIVA- und ACD-Stenose verlegt .
Bei der durchgeführten PET zeigte sich eine Minderperfusion anteroseptal und inferolateral , die Vitalität des gesamten
Myokards war jedoch erhalten , so daß wir uns zu einer Dilatation und Stenteinlage - sowohl der ACD als auch des
RIVA - entschlossen und diese am <date>18. und 19.08.1997</date> erfolgreich durchführten .

Wegen der Stentimplantation muß für die nächsten <date-relative>sechs Wochen</date-relative> die Gabe von
<branddrug>Tyklid</branddrug> erfolgen .
Dabei müssen im Abstand von zwei Wochen Blutbildkontrollen durchgeführt werden bis
<date-relative>zwei Wochen</date-relative>
nach der letzten Verabreichung .
Wir können <patient>Herr Stefanski</patient> jetzt beschwerdefrei in Ihre Betreuung zurückverlegen und danken für die
prompte Übernahme .

Empfohlene Medikation
<branddrug>Aponal</branddrug> 50
<chemicals>Ramipril</chemicals> <mass>5 mg</mass> (z. B.
<branddrug>Delix</branddrug> 5)
<branddrug>Centrum</branddrug> Tbl.

Wir danken Ihnen für die freundliche Überweisung und stehen für Rückfragen jederzeit gerne zur Verfügung .

Mit freundlichen , kollegialen Grüßen

<medical>PD Dr. med. K. Garbade</medical> <medical>Y. Erzberger</medical>
Chefarzt Assistenzärztin

Abbildung 4: Beispiel eines automatisch generierten Arztbriefes

Der spätere Einsatz der Klassifizierer macht eine Konvertierung in das IOB-Format[28] notwendig. Dazu erfolgte zuerst eine Konvertierung in ein Textformat ohne Formatierungen. Als nächstes mussten die Sätze so zerlegt werden, dass in jeder Zeile nur ein Token gefolgt von der jeweiligen Bezeichnung der Entitätsklasse steht, z.B. „Müller .....PATIENT“. Handelte es sich um ein Wort, welches keiner Klassifizierung bedarf, wurde das Tag „O“ gesetzt. Das Ende eines Satzes wurde durch eine Leerzeile im IOB-Dokument gekennzeichnet. Da in den ersten Abschnitten eines jeden Arztbriefes bis zum Beginn des Freitextes (Briefkopf, Adressat, Datum, Betreff, etc.) viele Informationen vorkamen, die lediglich nur durch einen Zeilenumbruch voneinander getrennt waren (z.B. Name, Straße, Ort in der Empfängeranschrift), war ein besonderes Verhalten des Konvertierungsskriptes erforderlich. Bis zum Schlüsselwort „#newline-off“ (siehe Abbildung 4) wurde nach jedem Zeilenumbruch des Quelldokumentes, also im Arztbrief, eine leere Zeile in die zu erstellende IOB-Datei eingefügt. Danach erfolgte das Einfügen einer leeren Zeile in die Zieldatei nur noch nach einem Satzzeichen im Quelldokument, welches das Ende eines Satzes anzeigt.

4.2.2 Zusammensetzung des Trainings- und Evaluationskorpus

Bei diesem Experiment kam zum einen das manuell annotierte FRAMED-Korpus [Wermter04], bestehend aus 3.486 Sätzen, zum anderen das Korpus fiktiver Arztbriefe, welches 164.459 Sätzen enthielt, zum Einsatz. Letzteres wurde automatisch mithilfe der beschriebenen Datenbank generiert. Die generierten Dateien wurden im Verhältnis 90:10 in ein Trainingsset von 148.014 Sätzen sowie einen Goldstandard von 16.445 Sätzen aufgeteilt.

[28] Inside Outside Beginning. In der Computerlinguistik verwendetes Dateiformat.

4.3 Ergebnisse

Für die Interpretation der Ergebnisse ist hervorzuheben, dass der Klassifizierer ausschließlich mit künstlich generierten Texten trainiert wurde. Getestet wurde er dagegen immer an zwei Textarten: Einerseits mit nach demselben Schema künstlich generierten Texten, andererseits mit authentischen, manuell annotierten Sätzen. Der Grund für die Verwendung fiktiver Dokumente liegt im geltenden Datenschutz. Es ist höchst kompliziert, wenn nicht gar unmöglich, echtes Datenmaterial in Form von Arztbriefen aus einer Klinik für Forschungszwecke zu bekommen.

Die Klassifizierer wurden zunächst mit 20.000 Sätzen aus dem automatisch generierten Korpus trainiert. Die anschließende Anwendung der Klassifizierer auf 3.500 Sätze, die ebenfalls automatisch generiert wurden, jedoch nicht zum Trainingsmaterial gehörten, erbrachte mit einem F-Measure von 99,04% eine sehr hohe Rate an korrekt durchgeführten Identifizierungen der betrachteten personenbezogenen Textbestandteile vom Typ Titel, Vorname, Nachname, Straße, Ort, Telefonnummer, Krankenhaus und Station bzw. Praxis (siehe Tabelle 6). Eine Erhöhung des Trainingsmaterials auf 150.000 Sätze konnte diesen Wert nur um 0,21% erhöhen. Analog wurden die trainierten Klassifizierer auch auf realen, manuell annotierten Arztbriefe angewendet, um eine Vorhersage darüber zu treffen, wie das trainierte System auf einem unbekannten, realistischen Korpus funktionieren würde. Diese Untersuchung ergab einen F-Measure von nur 26,6% bzw. 24,4% nach einer Erhöhung des Trainingsmaterials auf 150.000 Sätze.

	evaluated: EXT (3.500)	evaluated: MAN (3.500)
trained: EXT (20.000)	**99,04%**	**26,6%**
trained: EXT (150.000)	**99,25%**	**24,4%**

Tabelle 6: F-Measure bei der Evaluierung eines Trainingskorpus. Die Zahl in den Klammern gibt die Anzahl der verwendeten Sätze an. MAN: realistisches, manuell annotiertes Datenmaterial. EXT: automatisch generiertes Datenmaterial

Die geringe Quote an Übereinstimmungen bei Anwendung der trainierten Klassifizierer auf einem realen Korpus (MAN) gab Anlass zu einer detaillierten Auswertung der unterschiedlichen Textbestandteile (siehe Tabellen 7 und 8). Zu erst wurde untersucht, wie viele der im Korpus insgesamt vorhandenen anonymisierungsrelevanten Textbestandteile erkannt wurden (Spalten 1 und 2; in der Tabelle als *entity* bezeichnet). Daraufhin erfolgte eine Analyse nach Klassen (z.B. *person*; Spalten 3 und 4) und schließlich nach Typen (z.B. *patient*; Spalten 5 und 6), jeweils getrennt nach Evaluation auf automatisch generierten Arztbriefen (EXT) und einem Korpus realer medizinischer Dokumente (MAN). Da der Typ *office* (Arztpraxis) weder im Trainings- noch im Evaluationskorpus vorkam, wurde er nicht weiter beachtet und nicht in die Tabelle aufgenommen.

Eine differenzierte Betrachtung korrekt identifizierter Textbestandteile nach Tokens (siehe Tabelle 7) und Tags (siehe Tabelle 8) stellt eine Verschärfung der Auswertung dar. Zu erst erfolgte eine Auswertung nach Tokens, also der Überprüfung, ob ein Token bei der Anwendung der trainierten Klassifizierer im Vergleich zum Goldstandard korrekt klassifiziert wurde. Häufig bestand ein Tag aus mehreren Tokens, wie z.B. „Herr Prof. Dr. Anton Bamberger“. Würde das Wort „Bamberger“ nicht als Name identifiziert werden und nur die ersten vier Tokens mit dem Tag einer Person versehen werden, wäre die Rate der korrekten Erkennung von 4/5 bei vier von fünf richtig erkannten Tokens die Vortäuschung eines guten Ergebnisses. Betracht man die Tatsache, dass jedoch das wichtigste Element des Tags, nämlich der Nachname, nicht als solcher erkannt wurde und folglich in der Folge nicht anonymisiert werden kann, war die gesamte Klassifizierung nicht hilfreich und damit eine Erkennungsquote von 80% falsch hoch.

<table>
<tr><td rowspan="12">entity</td><td>EXT</td><td rowspan="6">person</td><td>EXT</td><td rowspan="2">patient</td><td>EXT</td></tr>
<tr><td rowspan="11">P: 3294/3299=0,998
R: 3294/3304=0,997
F: 0,998</td><td rowspan="2">P: 1324/1324=1,0
R: 1324/1328=0,997
F: 0,998</td><td>P: 355/355=1,000
R: 355/259=0,989
F: 0,994</td></tr>
<tr><td>MAN</td><td rowspan="2">P: 48/186=0,258
R: 48/80=0,600
F: 0,361</td></tr>
<tr><td>MAN</td><td rowspan="2">medical</td><td>EXT</td></tr>
<tr><td rowspan="2">P: 75/231=0,325
R: 75/117=0,641
F: 0,431</td><td>P: 969/969=1,000
R: 969/969=1,000
F: 1,000</td></tr>
<tr><td>MAN</td></tr>
<tr><td>P: 22/45=0,489
R: 22/37=0,595
F: 0,537</td></tr>
<tr><td rowspan="6">address</td><td>EXT</td><td rowspan="3">street</td><td>EXT</td></tr>
<tr><td rowspan="3">P: 1250/1253=0,998
R: 1250/1264=0,989
F: 0,993</td><td>P: 302/305=0,990
R: 302/306=0,987
F: 0,989</td></tr>
<tr><td>MAN</td></tr>
<tr><td>P: 2/61=0,033
R: 2/2=1,000
F: 0,063</td></tr>
<tr><td rowspan="2">city</td><td>EXT</td></tr>
<tr><td>P: 443/443=1,000
R: 443/453=0,978
F: 0,989</td></tr>
</table>

<table>
<tr><td rowspan="14">entity</td><td>MAN</td><td rowspan="5">address</td><td>MAN</td><td rowspan="3"></td><td>MAN</td></tr>
<tr><td rowspan="13">P: 151/541=0,279
R: 151/240=0,629
F: 0,387</td><td rowspan="4">P: 8/121=0,066
R: 8/23=0,348
F: 0,111</td><td>P: 4/51=0,078
R: 4/18=0,222
F: 0,116</td></tr>
<tr><td rowspan="3">phone</td><td>EXT</td></tr>
<tr><td>P: 505/505=1,000
R: 505/505=1,000
F: 1,000</td></tr>
<tr><td>MAN</td></tr>
<tr><td>P: 2/9=0,222
R: 2/3=0,667
F: 0,333</td></tr>
<tr><td rowspan="8">organisation</td><td>EXT</td><td rowspan="3">hospital</td><td>EXT</td></tr>
<tr><td rowspan="3">P: 711/722=0,985
R: 711/712=0,999
F: 0,992</td><td>P: 155/163=0,951
R: 155/156=0,994
F: 0,972</td></tr>
<tr><td>MAN</td></tr>
<tr><td>P: 12/59=0,203
R: 12/28=0,429
F: 0,276</td></tr>
<tr><td>MAN</td><td rowspan="3">dept</td><td>EXT</td></tr>
<tr><td rowspan="3">P: 62/189=0,328
R: 62/100=0,620
F: 0,429</td><td>P: 556/559=0,995
R: 556/556=1,000
F: 0,997</td></tr>
<tr><td>MAN</td></tr>
<tr><td>P: 43/130=0,331
R: 43/72=0,597
F: 0,426</td></tr>
</table>

Tabelle 7: Auswertung nach Tokens. MAN: realistisches, manuell annotiertes Datenmaterial. EXT: automatisch generiertes Datenmaterial. P: Precision. R: Recall. F: F-Measure

<table>
<tr><td rowspan="12">entity</td><td>EXT</td><td rowspan="6">person</td><td>EXT</td><td rowspan="3">patient</td><td>EXT</td></tr>
<tr><td rowspan="11">P: 1203/1208=0,996
R: 1203/1215=0,990
F: 0,993</td><td rowspan="2">P: 387/387=1,000
R: 387/391=0,990
F: 0,995</td><td>P: 182/182=1,000
R: 182/186=0,978
F: 0,989</td></tr>
<tr><td>MAN</td></tr>
<tr><td rowspan="3">patient</td><td>P: 24/90=0,267
R: 24/40=0,600
F: 0,369</td></tr>
<tr><td>MAN</td><td rowspan="3">medical</td><td>EXT</td></tr>
<tr><td>P: 30/104=0,288
R: 30/53=0,566
F: 0,382</td><td>P: 205/205=1,000
R: 205/205=1,000
F: 1,000</td></tr>
<tr><td>MAN</td></tr>
<tr><td>P: 4/14=0,286
R: 4/13=0,308
F: 0,296</td></tr>
<tr><td rowspan="5">address</td><td>EXT</td><td rowspan="3">street</td><td>EXT</td></tr>
<tr><td rowspan="4">P: 609/612=0,995
R: 609/619=0,984
F: 0,989</td><td>P: 127/130=0,977
R: 127/131=0,969
F: 0,973</td></tr>
<tr><td>MAN</td></tr>
<tr><td>P: 1/27=0,037
R: 1/1=1,000
F: 0,071</td></tr>
<tr><td rowspan="2">city</td><td>EXT</td></tr>
<tr><td>P: 257/257=1,000
R: 257/263=0,977
F: 0,988</td></tr>
</table>

<table>
<tr><td rowspan="12">entity</td><td>MAN</td><td rowspan="4">address</td><td>MAN</td><td rowspan="2"></td><td>MAN</td></tr>
<tr><td rowspan="11">P: 56/261=0,215
R: 56/123=0,455
F: 0.292</td><td rowspan="3">P: 6/73=0,082
R: 6/20=0,300
F: 0,129</td><td>P: 3/37=0,081
R: 3/16=0,188
F: 0,113</td></tr>
<tr><td rowspan="3">phone</td><td>EXT</td></tr>
<tr><td>P: 235/235=1,000
R: 235/235=1,000
F: 1,000</td></tr>
<tr><td rowspan="7">organisation</td><td>MAN</td></tr>
<tr><td>P: 2/9=0,222
R: 2/3=0,667
F: 0,333</td></tr>
<tr><td>EXT</td><td rowspan="3">hospital</td><td>EXT</td></tr>
<tr><td rowspan="2">P: 333/341=0,977
R: 333/334=0,997
F: 0,987</td><td>P: 71/76=0,934
R: 71/72=0,986
F: 0,959</td></tr>
<tr><td>MAN</td></tr>
<tr><td>P: 3/26=0,115
R: 3/16=0,188
F: 0,143</td></tr>
<tr><td>MAN</td><td rowspan="3">dept</td><td>EXT</td></tr>
<tr><td>P: 18/84=0,214
R: 18/56=0,321
F: 0,257</td><td>P: 262/265=0,989
R: 262/262=1,000
F: 0,994</td></tr>
<tr><td>MAN</td></tr>
<tr><td>P: 15/58=0,259
R: 15/40=0,375
F: 0,306</td></tr>
</table>

Tabelle 8: Auswertung nach Tags. MAN: realistisches, manuell annotiertes Datenmaterial. EXT: automatisch generiertes Datenmaterial. P: Precision. R: Recall. F: F-Measure

Textbestandteil	tatsächlich vorhanden	davon gefunden	Recall
Patienten- und Ärztenamen	117	75	61,4%
Patientennamen	80	48	60,0%

Tabelle 9: Erkennung von Textbestandteilen als anonymisierungsrelevant unabhängig von der Richtigkeit des zugewiesenen Typs

Da unter den anonymisierungsrelevanten Textbestandteilen die Erkennung von Namen, sei es von Patienten oder Ärzten, die größte Bedeutung hat, wurden einige zusätzliche Auswertungen durchgeführt. Die Erkennung eines Textbestandteils als anonymisierungsrelevant ist wesentlich wichtiger als die Identifizierung des Typs. Wenn beispielsweise der Name eines Patienten als Straße klassifiziert wurde und in der Folge unkenntlich gemacht wird, dann ist dies bei weitem weniger schlimm, als wenn der Patientenname gar nicht erkannt würde. Deshalb wurde untersucht, wie viele Namen als anonymisierungsrelevant erkannt wurden, unabhängig von der Richtigkeit des zugewiesen Typs. Der dazu berechnete Recall entsprach allerdings den bereits vorhanden Ergebnissen aus der Analyse der einzelnen Typen und Klassen (siehe Tabelle 9).

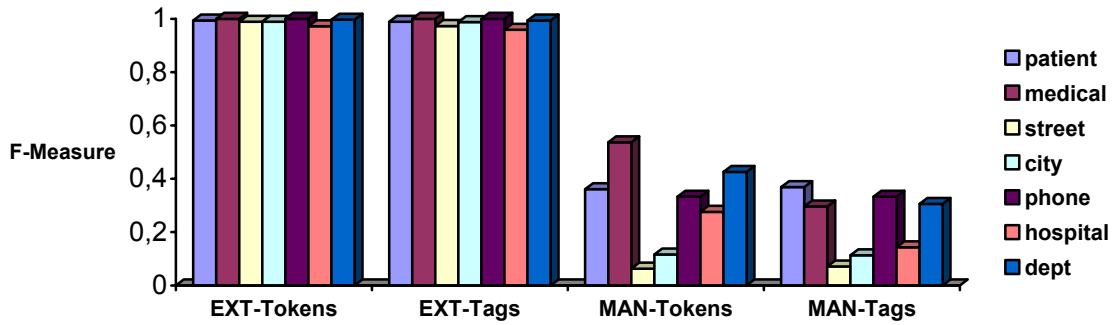

Abbildung 5: Graphische Darstellung des F-Measure als zusammenfassendes Maß von Precision und Recall

5 Diskussion

Die hier vorgestellten Experimente sind die ersten ihrer Art, die sich mit dem effizienten Trainieren von Klassifzierern im Hinblick auf eine Anonymisierung deutschsprachiger medizinischer Dokumente beschäftigen. Ein zentrales Studienergebnis ist eine mögliche Reduktion des Annotationsaufwandes um ca. 70% durch Anwendung einer Active Learning - Selektionsstrategie. Der Ansatz, ein Modell durch künstlich generierte Daten zu trainieren, hat sich als schwierig dargestellt und bedarf weiterer Arbeiten.

Damit ein medizinisches Dokument als anonymisiert betracht werden kann, müssen alle relevanten personenbezogenen Textbestandteile ausnahmslos als solche erkannt worden und entfernt sein. Wenn in der Folge von ausreichender Anonymisierung die Rede ist, handelt es sich immer um eine komplette Entfernung personenbezogener Informationen.

Auch die herkömmliche, höchst arbeitsaufwendige Unkenntlichmachung personenbezogener Daten durch eine reale Person, die in einem medizinischen Text diese personenbezogenen Textbestandteile streichen soll, erreicht die soeben definierte Erfolgsquote von 100% nicht. So wurden in einem solchen Experiment bis zu sechs Prozent der sensiblen Informationen von einem Menschen nicht als solche markiert und entsprechend anonymisiert [Sweeney96].

5.1 Interpretation der durchgeführten Experimente

Die in Experiment I angewandte Active Learning - Selektionsstrategie hat gezeigt, dass nur ca. 30% der vorhanden Textmenge annotiert werden müssen, um eine ähnliche Performanz wie bei der Annotation der kompletten Textdaten zu erreichen. Die zu diesem Zweck betrachteten Datums- und Zeitangaben zeichneten sich dadurch aus, dass sie aufgrund ihrer unterschiedlichen

Erscheinungsformen automatisch schwer zu erfassen waren. Trotzdem konnte eine signifikante Reduktion des Annotationsaufwandes gezeigt werden.

Es stellt sich nun die Frage, ob ähnliche Ergebnisse bei der Übertragung auf andere Informationsarten als Datums- und Zeitangaben erzielt werden können. In diesem Zusammenhang wäre es denkbar, mit Diagnosen, Dosierungen sowie Medikamentnamen und Wirkstoffbezeichnungen zu experimentieren. Dies wäre in besonderem Maße zur automatischen Informationsextraktion in klinischen Texten und bei Textrecherchen nach semantischen Gesichtspunkten von Interesse. Eine automatische Analyse von Arztbriefen könnte beispielsweise bei statistischen Fragestellungen zur Anwendung kommen, um den Einsatz von Medikamenten bei bestimmten Diagnosen zu untersuchen.

Das Experiment I hatte den Charakter einer Machbarkeitsstudie. Es diente als Vorbereitungsschritt um der Frage nachzugehen, ob eine Anonymisierung von klinischen Patientendaten realistisch ist. Daraufhin wurde in Experiment II untersucht, ob eine automatische Generierung von fiktiven Arztbriefen möglich ist, um Klassifizierer effektiv zu trainieren. Die Ergebnisse zeigten, dass nach einem Training mit einem Korpus von 20.000 Sätzen bei der Evaluation auf fiktiven Texten ein F-Measure von 99,04% erreicht werden konnte, der sich auch bei einer Erhöhung des Trainingskorpus auf 150.000 Sätze nur marginal erhöhte. Gleichermaßen bedeutete dieses effizient scheinende Ergebnis aber auch, dass ein von hundert relevanten personenbezogenen Textbestandteilen in dem generierten Korpus nicht als solches erkannt wurde. Das Ergebnis ist aber dennoch wesentlich besser als bei einer manuellen Anonymisierung.

Bei der Evaluation auf dem fiktiven Trainingskorpus fiel auf, dass die Erkennungsraten der unterschiedlichen personenbezogenen Textbestandteile recht ähnlich waren. Durchweg lagen sie bei über 96%. Am schwierigsten war dabei die Identifizierung von Krankenhausnamen und Straßen. Beide Textbestandteile bestanden häufig aus mehreren Teilen, so dass eine Erkennung

aller Tokens, die zu einem Tag gehören, erschwert war (siehe Tabelle 10). Die zuverlässigste Markierung erfolgte bei Telefonnummern und Arztnamen. Das lässt sich dadurch erklären, dass Arztnamen in aller Häufigkeit mit einem Titel (Prof., Dr.) als so genannte Identitätsmarker begonnen. Bei Ärzten ohne akademischen Titel einerseits und promovierten Patienten andererseits kam es zu falsch negativen bzw. falsch positiven Ergebnissen. Telefonnummern waren durch ihre charakteristische Zusammensetzung aus Ziffern leicht zu erkennen.

Kreiskrankenhaus Neustadt an der Elbe	Joschka Fischer Straße
Kurklinik Titisee Neustadt	Am Stühm Süd

Tabelle 10: Zusammengesetzte Krankenhaus- und Straßennamen als Beispiele für schwierig zu fassende personenbezogene Textbestandteile

Bei dieser Evaluation auf dem fiktiven Korpus wurde deutlich, dass trotz des Trainings mit einem fiktiven Goldstandard kein F-Measure von 100% erzielt werden konnte. Dies erlaubt den Schluss, dass die Ergebnisse auf einem realen Korpus schlechter sein müssen.

Evident ist die Frage, warum die Evaluation auf einem realen Korpus wesentlich schlechter als erwartet ausfiel. Eine wichtige Ursache liegt dabei in der Tatsache, dass extrem wenig und sehr heterogenes Datenmaterial für die Evaluation zur Verfügung stand. Das Korpus bestand aus einer Mischung von Arztbriefen ohne Briefkopf[29], Pathologiebefunden und OP-Berichten. Das bedeutet, dass die Verteilung der personenbezogenen Textbestandteile im Evaluationskorpus und im generierten Korpus bei weitem nicht dieselben waren. Nur ein Dokument im Evaluationskorpus enthielt einen Arztbrief samt Briefkopf und Adressangaben. Die darin enthaltene Straße („Meierstr. 8“) entsprach einem Tag und zwei Tokens. Diese wurde auch korrekt von unserem

[29] In Briefköpfen kommen häufig auf engstem Raum viele personenbezogene Informationen in relativ geordneter Struktur vor. Diese Textbestandteile sind in der Regel durch Klassifizierer

System erkannt (Recall = 100%, allerdings mit großem Konfidenzintervall). Die Precision von 3,7% erlaubt jedoch die Aussage, dass zu viele Tokens im Evaluationskorpus als Straßen erkannt wurden.

Bei der Behandlung von Eigennamen fiel ebenfalls auf, dass die Rate der fälschlicherweise als relevant erkannten personenbezogenen Textbestandteile recht hoch lag. Häufig waren davon Tokens betroffen, die aus dem medizinischen Bereich stammen, aber nicht der in der Medizin verwendeten lateinisch-griechischen Terminologie entsprechen (siehe Tabelle 11).

Redon - Drainage	Streak - Gonaden Beidseits
Grand mal - Anfallsleiden	Cushingoider Habitus

Tabelle 11: Falsch richtig erkannte personenbezogene Textbestandteile in der Kategorie *patient* bei Evaluation auf dem realen Korpus

In der Gruppe der Telefonnummern haben Zahlenkombinationen Probleme bereitet, die Dosisangaben entsprachen (z.B. 1-0-1). In den restlichen Kategorien haben sich keine Muster finden lassen, die die Falsch-Positivität erklären könnten.

5.2 Einordnung in derzeit diskutierte Modelle

Bei dem in diesen Studien untersuchten Thema, dem effizienten Training von Klassifizierern, handelt es sich um ein Gebiet der Fragestellung, wie anonymisierungsrelevante Informationen in medizinischen Dokumenten zu finden sind. Deshalb erfolgt im Folgenden eine Betrachtung von Anonymisierungsmodellen.

---

leicht zu erkennen.

Das von Sweeney realisierte Scrub System sucht in englischen Dokumenten nach Zeichenketten, die Rückschlüsse auf die Identität von Patienten möglich machen könnten [Sweeney96]. Diese Studie gehört zu den ersten, die zu diesem Thema durchgeführt wurden. Durch die Anwendung von Methoden, die nach der Struktur von persönlichen Informationen suchen (z.B. Identity Markers wie „Herr/Frau“ für Personennamen) und durch die Kenntnis häufig vorkommender Namen sowie Daten aus den Krankenakten, konnte nach eigenen Angaben eine Trefferquote von 99-100% erzielt werden. Im Vergleich dazu erkennen Anwendungen, die lediglich nach bestimmten in einer Liste vorgegebenen Wörtern suchen, nur 30-60% der relevanten Referenzen. Sweeney entwickelte ein Modell, dass das Vorgehen von Menschen bei der Suche der relevanten Information imitieren sollte. Dazu gehörte unter anderem das Anlegen von Listen, welche häufig vorkommende Namen enthielten. Dies simulierte die menschliche Kenntnis, Wörter wie „Smith“ und „Jones“ eindeutig als Nachnamen zu erkennen. Für jeden anonymisierungsrelevanten Textbestandteil (z.B. Name, Straße) wurden dann Algorithmen zur Erkennung entwickelt. Nach Sweeney hat es sich als sinnvoll erwiesen, einmal identifizierte personenbezogene Textbestandteile allen anderen Algorithmen zugänglich zu machen. Wurde beispielsweise ein vollständiger Patientenname erkannt und taucht in demselben Dokument der gleiche Vor- oder Nachname – evtl. unabhängig von einander – ein weiteres Mal auf, so kann dieses Wissen bei der Erkennung nützlich sein. Zusammenfassend zeigt die Studie, dass durch die Anwendung von Listen und Strukturanalysen von Tokens[30] eine akzeptable Anonymisierung eines beschränkten Datenguts erfolgen kann.

Mit 99-100% erscheint die von Sweeney angegebene Trefferquote als sehr hoch. Sweeney nutzt in ihrer Arbeit lediglich Patientenmaterial aus einer Abteilung für Kinderheilkunde, der ein hoher Grad an Homogenität unterstellt werden kann. Im allgemeinen Einsatz sollte eine Anonymisierungsroutine aber nicht auf ein Fachgebiet und einen Dokumententyp beschränkt sein, sondern

[30] Bezeichnung für Zeichenkette in der Linguistik

möglichst universell auf unterschiedlichstes Material angewendet werden können. Hierbei handelt es sich um einen ernstzunehmenden Schwierigkeitsfaktor. Selbst standarisierte Arztbriefe aus Kliniken unterschiedlicher Fachrichtungen differieren stark voneinander. Ferner gehen aus Sweeneys Studie keine quantitativen Angaben über Trainings- und Testdatensätze hervor, die für die Beurteilung einer Trefferquote von Bedeutung wären.

Ohrn *et al.* stellen ein Anonymisierungssystem vor, dass aufgrund von booleschen Folgerungen ein medizinisches Dokument anonymisieren kann [Ohrn99]. Damit ist gemeint, dass vom Modell erkannte Zusammenhänge in Form von Aussagen repräsentiert werden, die entweder wahr oder falsch sein können. Mithilfe der Logik lassen sich daraus Folgerungen schließen. Abhängig von der gewünschten „Anonymisierungstiefe“ kann im vorgestellten Modell festgelegt werden, wie hoch der Grad an entfernten Textbestandteilen sein soll.

In [Ruch00] kommen ein semantisches Lexikon für medizinische Daten und ein Toolkit für *word-sense* (d.h. nach dem Wortsinn urteilend) und *morpho-syntactic* (d.h. von morphologischen Gesichtspunkten ausgehend) Tagging zum Einsatz. Beide Elemente stammen aus dem MEDTAG-Framework [Ruch99]. Bei der Verwendung von Taggern wird nach Identitätsmarkern gesucht, die bestimmte personenbezogene Textbestandteile einleiten. Kann nicht eindeutig anhand eines Markers über die Art eines anonymisierungsrelevanten Textbestandteils entschieden werden, können die word-sense- und morphosyntaktischen Komponenten helfen, diese Ambiguitäten aufzulösen. Zur Extraktion und Entfernung der ausgewählten personenbezogenen Textbestandteile entschieden sich die Autoren für formal recursive transition networks.[31] Die Arbeit verifiziert folgende methodische Überlegungen: Zum einen kann Syntax helfen, die Bedeutung von Wörtern zu unterscheiden, welche unterschiedlichen syntaktischen Kategorien angehören. Des weiteren können syntaktische und semantische Ambiguitäten durch einfache Tagger aufgelöst werden, und

schließlich kann die Extraktion der Informationen ebenfalls tag-unterstützt erfolgen. Im konkreten Fall haben die Autoren etwa 30 Regeln manuell erarbeitet, um mit den verschiedenen Taggern (seriell angewendet) bei Ambiguitäten zu ermitteln, ob es sich um eine anonymisierungsrelevante Entität handelt oder nicht. Schließlich konnten durch den Einsatz dieser NLP-Mechanismen 96,8% der fraglichen Tokens korrekt entfernt werden.

Gupta *et al.* wenden eine Vielzahl von Mechanismen an, um ihrem System eine hohe Performanz zu ermöglich [Gupta04]. So kommen ein regelbasierter Tagger, Wörterbücher, Pattern-Matching-Algorithmus[32] sowie das *Unified Medical Language System* UMLS[33] zum Einsatz. Die Kombination all dieser Methoden erlaubte die Fertigstellung eines Prototyps, der relativ zuverlässig die wichtigsten personenbezogenen Informationen eines Dokuments entfernen kann.

Anlässlich der AMIA[34] 2006 fand ein Workshop über maschinelle Sprachverarbeitung im medizinischen Bereich statt.[35] Im Rahmen dieses Symposiums erfolgte eine Ausschreibung zur Entwicklung eines Modells, welches persönliche Referenzen in medizinischen Dokumenten (*Private Health*

[31] Für die theoretischen Grundlagen sei auf [Gazdar89] verwiesen.

[32] Algorithmus, der nach bestimmten Mustern sucht, die für das auftreten der gesuchten Information typisch sind.

[33] Ein seit 1989 von der US National Library of Medicine unterhaltenes System, welches eine Vielzahl medizinischer Terminologiesysteme auf gemeinsame semantische Deskriptoren abbildet (Metathesaursus) und mit einem umfangreichen Lexikon an Medizintermen (Spezialist Lexicon) verknüpft ist. Für ein Modell zur semantischen Klassifikation unbekannter Wörter mithilfe von UMLS sei auf [Campbell99] verwiesen.

[34] Jahreskonferenz der *American Medical Informatics Association*, http://www.amia.org

[35] Fall Symposium Workshop on Medical Natural Language Processing, https://www.i2b2.org/de-id/

Information, PHI) entfernen sollte.[36] Eine Evaluation der eingesandten Beiträge findet sich in [Uzuner07].

Zu den Beiträgen gehört das Modell von Aramaki *et al.* [Aramaki06]. Auch sie nutzen CRF zur Identifizierung relevanter Textbestandteile. Das Besondere an der Arbeit ist die Beobachtung, dass die meisten sensiblen Textbestandteile in Arztbriefen häufig am Anfang und am Ende des Dokumentes zu finden sind. Außerdem erscheint die Länge derjenigen Sätze kürzer, die relevante Informationen enthalten.

Guillen stellt ein regelbasiertes Modell vor, das globale (z.B. Satzposition), lokale (z.B. besondere Zeichen) und syntaktische Features zur Erkennung von sensiblen Informationen nutzt [Guillen06]. Reguläre Ausdrücke werden zum Finden von Datumsangaben und anderen Ziffernkombinationen (z.B. Patientennummern) genutzt; lexikalische Informationen dienen dem Auffinden bestimmte Begriffe (z.B. lässt „discharge summary name: “ einen Namen als nächstes Token erwarten).

Guo *et al.* nutzen *support vector machines*[37] und das GATE-System[38] für zwei vorgelegte Modelle [Guo06]. Durch den Vergleich beider Beiträge zeigen die Autoren, dass bei der Erkennung von Patientennamen der Zusammenhang, in dem sich das Token befindet, eine größere Rolle spielt als die Information, die durch ein reines System zur Eigennamenerkennung gewonnen wird.

---

[36] Die Resultate des Workshops lagen erst nach Abschluss der Experimentierphasen der hier durchgeführten Studien vor und konnten deshalb im gewählten Studiendesign nicht berücksichtigt werden.

[37] Für die theoretischen Grundlagen sei auf [Görz03] verwiesen.

[38] General Architecture for Text Engineering. Vergleiche [Natural Language Processing Group08].

Wellner *et al.* und Szarvas *et al.* haben mit ihren vorgeschlagenen Ansätzen die besten Ergebnisse erzielt. Wellners Arbeit beruht auf der Anwendung zweier bereits existierender Toolkits für die Erkennung von Eigennamen [Wellner07]. In nur vier Stunden gelang es, eines der beiden Tools an die gestellten Anforderungen anzupassen. Durch eine spätere Implementierung von Attribut-Analysen und eines Lexikons für die Feinabstimmung des Systems wurde ein F-Measure-Wert von 0,9736 erzielt. Auch hier kamen CRF zum Einsatz.

Szarvas *et al.* nutzt Entscheidungsbäume mit lokalen Attributen und Wörterbüchern [Szarvas07]. In einem ersten Schritt werden alle Tags markiert, deren Bedeutung aus der Struktur des Textes entnommen werden kann. Als nächstes werden diese Fundstellen dafür genutzt, weitere relevante anonymisierungsrelevante Textbestandteile im Freitext ausfindig zu machen. Das Modell arbeitet mit Inhaltsbeziehungen, die mithilfe von lexikalischen Triggern erstellt werden. Diese Trigger werden entsprechend der Stärke ihrer Affinität zu den einzelnen Entitätsklassen geordnet. Um Konsistenz innerhalb der Entitätsklassen zu gewährleisten, werden alle verteilten Labels nachbearbeitet. Dabei wird jedes Vorkommen eines zusammengesetzten Tokens mit dem Label der am längsten erkannten und passenden Tokenfolge versehen. Der erreichte F-Measure-Wert beträgt 0,9975.

Sponsor der hier diskutierten Studien der *Deidentification Challenge* ist das vom amerikanischen National Institutes of Health finanzierte Projekt *Informatics for Integrating Biology & the Bedside* (i2b2). Dieses hat eine lauffähige Demonstration einer De-Identifikationsanwendung zur Verfügung gestellt[39], die auf den Arbeiten von Sibanda *et al.* beruhen [Sibanda06a]. Das Modell arbeitet mit einem Korpus bestehend aus 340.000 Wörtern aus Arztbriefen. Aus [Sibanda06b] geht hervor, dass eine Kategorisierung der Semantik hilfreich für das Textverständnis. Dazu kann eine effektive Analyse der Syntax einen Beitrag leisten. Die vorgestellte statistische Routine ermöglicht die

Identifizierung von acht semantischen Kategorien. In den meisten dieser Kategorien werden F-Measures von über 90% erreicht.

Beckwith *et al.* stellen eine Lösung zur Anonymisierung von Pathologiebefunden vor [Beckwith06], die auf drei Teilen beruht. Zu erst wird nach patientenbezogenen Textbestandteilen mit expliziten Angaben (z.B. Name, Geburtsdatum, etc.) gesucht. In einem nächsten Schritt werden bestimmten Muster gesucht, die sich häufig in Verbindung mit persönlichen Informationen finden. Auf diese Weise werden Adressen, Arztnamen und ähnliches eliminiert. Letztlich findet ein Abgleich von Wörtern mit einer Datenbank aus Eigennamen und Orten statt. 98,3% aller relevanten Textbestandteile konnten auf diese Weise entfernt werden.

Keines der bisher erwähnten Verfahren kann als absolut zuverlässig bewertet werden. Die Gefahr, die Identität eines Patienten, dessen Daten anonymisiert wurden, herauszufinden (so genanntes *reverse scrubbing*), ist besonders bei außergewöhnlichen Fällen möglich [Sweeney96]. Dies können beispielsweise besonders seltene Krankheiten oder auch der Beruf bzw. die Stellung des Patienten sein, wenn dies, auch nur indirekt, im Dokument erwähnt ist. Dreiseitl *et al.* haben gezeigt, dass durch Anwendung komplexer Algorithmen Informationen zu einzelnen Personen aus anonymisierten Datenmaterial gewonnen werden kann [Dreiseitl01].

In diesem Zusammenhang stellt auch die Kombination verschiedener Datenquellen eine Gefahr dar. Sweeney konnte durch Kombination von zwei verfügbaren Datenbeständen zeigen, dass sich scheinbar anonymisierte Daten unter Umständen realen Personen zuordnen lassen [Sweeney02]. Die erste Datei ihres Experiments stammte von einem amerikanischen Krankenversicherer, der alle vorkommenden Eigennamen anonymisiert hatte. Die zweite Datei war das offizielle Wählerverzeichnis des Bundesstaates. Durch

[39] https://www.i2b2.org/de-id/demo.php

Verknüpfung von Informationen, wie Postleitzahl, Geburtsdatum und Geschlecht, die in beiden Dateien unanonymisiert vorkamen, konnten einige Patienten identifiziert werden.

5.3 Schwierigkeiten eines Anonymisierungssystems

Es sind zwei Arten von Schwierigkeiten zu unterscheiden. Zum einen geht es um konkrete, spezifische Textbestandteile, die in medizinischen Dokumenten auftauchen und von einem Anonymisierungssystem schwierig zu verarbeiten sind. Ferner gibt es allgemeine Schwierigkeiten, die sich bei dem Versuchsdesign ergeben und die Evaluierung mit Echtdaten erschweren.

Zu den konkreten Schwierigkeiten zählt die Vielzahl von Ambiguitäten. Beispiele sind Patientennamen, die Körperteilen („Frau Iris Leber“) und Krankheiten entsprechen („Bei Herrn Osler wurde ein Morbus Osler diagnostiziert“). Hier sind weitere Attribute zur Auflösung der Doppeldeutigkeiten erforderlich. Beispielsweise müsste in diesem Fall der Identitätsmarker „Herr“ dazu genutzt werden, diese Ambiguität aufzulösen.

Die Einbeziehung eines Wörterbuchs leistet einen entscheidenden Beitrag zur korrekten Identifizierung von Informationen [Ruch00]. In der sich schnell entwickelnden Sprache der Medizin wird es aber nie möglich sein, ein Wörterbuch zu entwickeln, das alle Elemente dieser Fachsprache enthält. Zu ungenau wäre eine mögliche Folgerung, bei einem Nichtvorhandensein eines Wortes im Wörterbuch daraus zu schließen, dass es sich um einen Eigennamen handelt.

Rechtschreibfehler, wie z.B. „Prfessor“, erschweren eine Identifizierung relevanter Textstellen ungemein, die die menschliche Intelligenz leicht korrigiert, von einem Algorithmus aber nicht ohne weiteres zu verarbeiten ist. Lexikalische Algorithmen können dabei einen großen Beitrag leisten [Ruch03].

Besondere biographische Details können die Anonymität wesentlich gefährden. Dazu zählen beispielsweise Referenzen wie „die Mutter von sieben Töchtern berichtete …“. Ab einem gewissen Alter des Patienten sollten besondere Mechanismen zum Tragen kommen. Patienten über 90 sollte daher in anonymisierten Dokumenten ein Standardalter von z.B. 92 Jahren zugewiesen werden.

Auch die Angabe des Berufs eines Patienten kann die Anonymität verletzen. Sollte es sich um einen kleinen Ort handeln, in dem der Patient lebt und sollte diese Angabe nicht anonymisiert worden sein, könnte relativ einfach auf die Identität der Person geschlossen werden, wie z.B. „der Schornsteinfeger aus Sankt Peter“. Je bekannter der Patient bzw. je höher die berufliche Stellung, desto größer ist die Gefahr einer Identifikation durch Dritte ("... dem Profirennradfahrer wurde an der Teilnahme an der diesjährigen Tour de France abgeraten", „dem Patienten wurde empfohlen, wegen der geplanten Chemotherapie nicht erneut für den Landtag zu kandidieren“). Derartige Referenzen sind von einem Computermodell außerordentlich schwierig zu erkennen.

Datumsangaben bedürfen einer besonderen Betrachtung. Für eine Verwendung in der Medizin sollten so wenige Änderungen wie möglich erfolgen. Für eine statistische Fragestellung kann es zum Beispiel interessant sein, eine repräsentative Anzahl von Arztbriefen darauf zu untersuchen, wie lange bei einer bestimmten Diagnose der durchschnittliche stationäre Aufenthalt beträgt. Eine mögliche Lösung wäre die Verschiebung der Daten nach einer Zufallszahl in einem vorher festgelegten Rahmen. Die Daten eines Arztbriefes über einen Patienten, der im Winter eine Fraktur nach einem Sturz beim Schlittschuh laufen erlitten hat, dürfen aber nicht in die Sommermonate verschoben werden. Eine Häufung von Brandverletzungen und Trommelfelltraumata in der Nacht vom 14. zum 15.1. legt den Schluss nahe, dass die Behandlungsdaten lediglich um zwei Wochen verschoben wurden. Auch das Patientenalter darf bei Frage-

stellungen zur Inzidenz und Prävalenz von Krankheiten nicht zu sehr verändert werden.

Zu den allgemeinen Schwierigkeiten zählen in erster Linie die ausgeprägte Heterogenität des Datenmaterials und die strengen Auflagen des Datenschutzes. Auch wenn Arztbriefe aus spezialisierten Abteilungen gewisse Ähnlichkeiten aufweisen, können sich Arztbriefe aus anderen Abteilungen strukturell und inhaltlich massiv unterscheiden. Es ist weiterhin anzunehmen, dass auch die verwendeten Fachausdrücke sehr unterschiedlich sind. Diese Heterogenität bei der Evaluation abzudecken ist unmöglich. Für die Testkorpora muss Datenmaterial aus dem bereits knappen Trainingskorpus entnommen werden, so dass nur wenig Material zur Evaluation zur Verfügung steht, welches für repräsentative Aussagen nicht geeignet ist.

Das Design dieser Studie war harten Bedingungen unterworfen. Bewusst wurde beispielsweise auf den Einsatz eines Lexikons verzichtet, da die bereits zitierten Studien den Nutzen einer solchen Referenz bewiesen haben. Die geringere Performanz des entwickelten Systems wurde also durchaus in Kauf genommen. Gleiches gilt für die Zusammensetzung des Korpus. Die meisten personenbezogenen Textbestandteile eines Arztbriefes befinden sich im standarisierten Briefkopf. Entsprechend stellt der Einsatz eines Korpus ohne dieses Informationskonglomerat eine erhebliche Erschwerung der Erkennungs-algorithmen dar.

Aus den Annotationsrichtlinien[40] geht hervor, dass 27 verschiedene Kategorien bei der Annotation angewendet wurden. Zwei dieser Kategorien, nämlich 4.1 date und 4.2 time wurden zusätzlich in relative und absolute Angaben sowie Zeitpunkte und Zeiträume unterteilt. Diese Vielfalt ist im Vergleich zu anderen Studien als hoch einzustufen. Je mehr Möglichkeiten bei der Klassifizierung der

[40] siehe Anhang 7.3

Tokens gegeben sind, je höher liegt zwangsläufig auch die Gefahr einer inkorrekten Annotation.

5.4 Anonymisierung vs. Pseudonymisierung

Eine Pseudonymisierung der sensiblen Daten bietet im Gegensatz zur Anonymisierung bestimmte Vorteile. Zum einen sind solche Dokumente wesentlich lesefreundlicher als kryptische Chiffrierungen, sofern als Pseudonym ein fiktiver Eigenname gewählt wurde. Zum anderen erlauben sie noch einen konkreten Bezug zu der realen Person. Bei der Vergabe von Pseudonymen, also dem Löschen konkreter Informationen zugunsten fiktiver Bezeichner, kann protokolliert werden, welches Pseudonym welchem Patientennamen in der Realität entspricht. Dies erlaubt eine Rückidentifizierung eines Falles, wenn Zugang zu dieser protokollierten Pseudonymvergabe besteht. Wenn umkehrt dieses Verzeichnis nicht verfügbar ist, lässt sich ein pseudonymisiertes Dokument genauso wenig einem Individuum zuordnen wie ein anonymisiertes [Pommerening05].

Die Rückidentifizierung ist zum Beispiel dann interessant, wenn Daten zu einem Fall aus verschiedenen Quellen oder von verschiedenen Zeitpunkten zusammengeführt werden [ebenda]. Getrennte Forschungsbereiche können so unterschiedlichen Fragestellungen nachgehen und trotzdem gefundene Ergebnisse mit Bezug zum selben Fall diskutieren. Wenn ein Pool langfristiger, qualitätsgesicherter Daten vorliegt, sind Langzeitbeobachtungen chronisch kranker Patienten, Erforschung von Spätfolgen und Auswirkungen auf die Lebensqualität nach aggressiven Therapien, etc. möglich [ebenda]. Auch für genealogische Studien wären medizinische Dokumente bei einer konsequenten Pseudonymisierung verwendbar [Ruch00].

Bei Versagen der Erkennung einer sensiblen Information hat die Pseudonymisierung mit Eigenamen einen entscheidenden Vorteil. Sollte ein

personenbezogenes Datum in einem Text nicht pseudonymisiert worden sein, so muss dies dem Leser nicht zwingend auffallen. Der Leser kann den nicht entsprechenden Namen durchaus für ein Pseudonym halten.

Pseudonyme bieten sich an, um eine autorisierte Zuordnung eines Dokumentes zu einem Patienten zu ermöglichen [Grätz05]. Bei Blutuntersuchungen in Arztpraxen ist dieses Verfahren seit langem üblich. Nach erfolgter Blutabnahme wird auf dem Anforderungszettel des Labors anstelle des Patientennamen lediglich eine Nummer eingetragen. Diese Nummer wird in der Arztpraxis zusammen mit dem Patientennamen in einer Liste festgehalten, so dass beim Eintreffen der Ergebnisse aus dem Labor die darauf vermerkte Nummer wieder dem Patienten zugeordnet werden kann. Auf diese Weise müssen keine persönlichen Daten aus der Arztpraxis an Dritte übermittelt werden.

Pommerening *et al.* haben ein generisches Datenschutzkonzept entwickelt, das in zwei Modellvarianten den Aufbau geeigneter Datenpools unter Verwendung von Pseudonymen beschreibt [Pommerening05]. Das Konzept wurde von Datenschutzbeauftragen genehmigt und dient als Grundlage individueller Konzepte für einzelne Netze. Auch Pommerening weist darauf hin, dass nicht ausgeschlossen werden kann, dass pseudonymisierte Daten einer realen Person zuzuordnen (Rückidentifizieren) sind. Bereits das Wissen über die Mitwirkung eines Patienten in einem bestimmten Forschungsnetz kann Aufschluss über seine Diagnose geben. Strikte Zugriffskontrollen sollen dieses Risiko minimieren.

5.5 Ausblick

Die gezeigte Reduktion des Annotationsaufwandes von Datum- und Zeitangaben in Experiment I zeigt die Effektivität von Active Learning zum Training probabilistischer Modelle zur maschinellen Sprachverarbeitung. Weiterführende Arbeiten sind nun nötig, um die Selektionsstrategie auf Eigennamen und andere personenbezogene Textbestandteile auszuweiten, die für eine Anonymisierung erforderlich sind.

Die Verwendung künstlich generierter Arztbriefe zum Training von Klassifizierern hat die Erwartungen nicht erfüllt. Ein Problem war dabei, dass das Evaluationskorpus recht klein und trotzdem sehr heterogen war. Es konnten dadurch keine repräsentativen Aussagen getroffen werden. Der diskutierte Workshop des *Deidentification Challenge* gibt einen Überblick über sinnvolle Attribute, die implementiert werden könnten [Uzuner07]. Die Kombination der Klassifizierer mit externen Attributen, wie zum Beispiel einem medizinischen Wörterbuch, könnte ebenfalls zu einer Verbesserung beitragen.

Die Diskussion des aktuellen Standes von automatischen Anonymisierern macht deutlich, dass mit heutigen Mitteln fast alle personenbezogenen Daten aus Dokumenten entfernt werden können. Mit F-Measures von über 99% hat sich eine Kombination verschiedener Attributen, darunter Orthographie, Frequenz, Satzbau und Kontext, als bewährt erwiesen [Szarvas07]. An derart hohe Erkennungsraten schließt sich die Frage einer Umsetzung in die Praxis an. Es bleibt zu untersuchen, unter welchen Bedingungen die gewonnen Daten unter Beachtung des Restrisikos der Re-Identifizierung genutzt werden können. Denkbar wäre in diesem Zusammenhang eine Beschränkung des Einsatzes in wissenschaftlichen Forschungseinrichtungen mit einer vorliegenden Versicherung, die Daten vor missbräuchlicher Nutzung zu schützen.

Beim Einsatz von Pseudonymen ist sicherzustellen, dass das Verzeichnis, das Informationen darüber enthält, welchem Pseudonym welcher Patientennname entspricht, besonders geschützt wird. Dieser Schutz sollte zum einen durch neueste Verschlüsselungstechniken, zum anderen durch räumliche Sicherheitsvorschriften (z.B. Aufbewahrung des Speichermediums in einem Tresor) gewährleistet werden.

Es stellt sich abschließend die Frage, ob die im Zusammenhang mit der Anonymisierung gewonnen Kenntnisse in anderen Bereichen der Namenserkennung in Medizindokumenten Anwendung finden könnte. Hierbei ist von Schwierigkeiten auszugehen, da die Heterogenität der Daten ein großes Problem darstellt. Es ist davon auszugehen, dass zwei Arztbriefe aus zwei verschiedenen Abteilungen derart verschieden sein können, dass sich positive Anonymisierungsergebnisse eines Arztbriefes mit einem System nicht mit dem anderen Arztbrief reproduzieren lassen. Umso größer wird der strukturelle und inhaltliche Unterschied ganz verschiedener Texte sein, die sich in der klinischen Medizin finden.

6 Zusammenfassung

Lehre und Forschung in der Medizin haben einen hohen Bedarf an patientenbezogenen Daten. Der Zugriff auf diese Daten ist aufgrund strenger Datenschutzauflagen nur sehr eingeschränkt möglich. Anonymisierungssysteme bieten die Möglichkeit, diese Barriere zu umgehen und trotzdem die Patienteninteressen uneingeschränkt zu wahren.

Für die Erkennung anonymisierungsrelevanter Textbestandteile wurden verschiedene Ansätze diskutiert. Die Manipulation der Daten setzt eine formalisierte Repräsentation voraus, die nach geeigneten Standards erfolgen muss. Die daraufhin erfolgende Erkennung relevanter personenbezogener Textbestandteile erfolgt durch Klassifizierer, die für diese Aufgabe trainiert sein müssen. Conditional Random Fields zeigen sich aufgrund ihrer Robustheit gegenüber Rechtschreibfehlern und anderen orthographischen Variationen für dieses Einsatzgebiet als sehr geeignet. Alle Methoden aus diesem Bereich benötigen für diese Problemstellung hochwertiges Trainingsmaterial, von denen maßgeblich die Performance abhängt. Die Schwierigkeit der Beschaffung von solchem bereits anonymisierten Material bzw. die hohen Kosten, die durch die Arbeit menschlicher Annotatoren entstehen, stellen einen Flaschenhals bei der Entwicklung von Anonymisierungssystem dar.

Es konnte gezeigt werden, dass der Annotationsaufwand durch Anwendung einer Active Learning-Selektionsstrategie um ca. 70% reduziert werden kann. Der Ansatz, ein Modell durch künstlich generierte Daten zu trainieren, hat sich als schwierig dargestellt und bedarf weiterer Arbeiten. Das anspruchsvolle Design der Studie (bewusster Verzicht auf regelbasierte Modelle und Hilfskomponenten wie Lexika sowie Ignorierung von Informationskonglomeraten in den standarisierten Bereichen eines Dokumentes) ist für eine Performanz verantwortlich, die für die Praxis inakzeptabel ist. Es wird die Tatsache betont, dass Gegenstand der vorliegenden Studie nicht die Entwicklung eines möglichst universell einsetzbaren Anonymisierungssystem sein sollte, sondern vielmehr den Fragestellungen nachgegangen wurde, wie einerseits die effiziente Bereitstellung von Trainingsmaterial für Klassifizierer erfolgen kann und andererseits welche Anforderungen an ein universell einsetzbares Anonymisierungssystem gestellt werden.

7 Anhang

7.1 Literaturverzeichnis

Aramaki E, Miyo K (2006) Automatic Deidentification by Using Sentence Features and Label Consistency. i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data

Bake C, Blobel B, Münch P (2004) Datenschutz und Datensicherheit im Gesundheits- und Sozialwesen. 1. Aufl, Datakontext, Frechen, S. 163

Baud R, Ruch P (2002) The future of Natural Language Processing for Biomedical Applications. Int J Med Inform 67 (1-3): S. 1-5

Beckwith BA, Mahaadevan R, Balis UJ, Kuo F (2006) Development and evaluation of an open source software tool for deidentification of pathology reports. BMC Med Inform Decis Mak 6 S. 12

Bellare M, Rogaway P. (2005) Introduction to Modern Cryptography. http://www-cse.ucsd.edu/~mihir/cse207/classnotes.html (13.03.2008)

Berner ES, Detmer DE, Simborg D (2005) Will the wave finally break? A brief view of the adoption of electronic medical records in the United States. J Am Med Inform Assoc 12 (1): S. 3-7

Campbell DA, Johnson SB (1999) A technique for semantic classification of unknown words using UMLS resources. Proc AMIA Symp S. 716-20

Carstensen K, Ebert C, Endriss C, Jekat S, Klabunde R, Langer H (2004) Computerlinguistik und Sprachtechnologie. 2., überarb. und erw. Aufl, Elsevier, München, S. XIV, 642

Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG (2001a) A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 34 (5): S. 301-10

Chapman WW, Fizman M, Chapman BE, Haug PJ (2001b) A comparison of classification algorithms to automatically identify chest X-ray reports that support pneumonia. J Biomed Inform 34 (1): S. 4-14

Cortes C, Vapnik V (2004) Support-Vector Networks. Machine Learning 20 (3): S. 273-297

Cunningham H (1996) Information Extraction - a User Guide. Proceedings of the 16th International Conference on Computational Linguistics (COLING) 1 S. 466-471

Dale R, Moisl H, Somers H (2000) Handbook of natural language processing. Dekker, New York, NY [u.a.], S. XVIII, 943

Dan S. (2004) Multi-Criteria-based Active Learning for Named Entity Recognition. http://www.coli.uni-saarland.de/~dshen/publications/MScThesis.pdf (13.03.2008)

de Bruijn B, Martin J (2002) Getting to the (c)ore of knowledge: mining biomedical literature. Int J Med Inform 67 (1-3): S. 7-18

Deutsches Ärzteblatt. (2008) cme (continuing medical education). http://www.aerzteblatt.de/cme/info.asp (13.08.2008)

Dierks C (2004) Pseudonymisierung statt Verschlüsselung: eine Alternative für Telemedizin-Projekte. Ärzte Zeitung, 01.09.2004

Dreiseitl S, Vinterbo S, Ohno-Machado L (2001) Disambiguation data: extracting information from anonymized sources. Proc AMIA Symp S. 144-8

Friedman C, Hripcsak G, Shagina L, Liu H (1999) Representing information in patient reports using natural language processing and the extensible markup language. J Am Med Inform Assoc 6 (1): S. 76-87

Friedman C, Shagina L, Lussier Y, Hripcsak G (2004) Automated encoding of clinical documents based on natural language processing. J Am Med Inform Assoc 11 (5): S. 392-402

Gazdar G, Mellish C (1989) Natural Language Processing in Prolog: An Introduction to Computational Linguistics. Repr. 1994, Addison-Wesley, Wokingham, S. XV, 504

Görz G (2003) Handbuch der Künstlichen Intelligenz. 4., korrigierte Aufl, Oldenbourg, München [u.a.], S. XIV, 1041

Grätz PGv (2005) Patientenakten im Netz: Datenschutz ist schon dann effektiv, wenn die Patientennamen ungenannt bleiben. Ärzte Zeitung, 22.03.2005

Grishman R, Sundheim B (1996) Message Understanding Conference - 6: A Brief History. Proceedings of the 16th International Conference on Computational Linguistics (COLING) I S. 466–471

Guillen R (2006) Automated De-Identification and Categorization of Medical Records. i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data

Guo Y, Gaizauskas R, Roberts I, Demetriou G, Hepple R (2006) Identifying Personal Health Information Using Support Vector Machines. i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data

Gupta D, Saul M, Gilbertson J (2004) Evaluation of a deidentification (De-Id) software engine to share pathology reports and clinical documents for research. Am J Clin Pathol 121 (2): S. 176-86

Heckl RW (1990) Der Arztbrief. 2., durchges. Aufl, Thieme, Stuttgart [u.a.], S. IX, 129

Hirschman L, Yeh A, Blaschke C, Valencia A (2005) Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC Bioinformatics 6 Suppl 1 S. S1

Krüger-Brand HE (2002) E-Learning in der Medizin: Vor dem Durchbruch. Deutsches Ärzteblatt (22/99): S. A-1491 / B-1270 / C-1193

Lafferty J, McCallum A, Pereira F. (2001) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. http://www.cis.upenn.edu/~pereira/papers/crf.pdf (13.03.2008)

Luger GF (2003) Künstliche Intelligenz. 4. Aufl., [Nachdr.], Pearson Studium, München, S. 892

Manning CD, Schütze H (2005) Foundations of statistical natural language processing. MIT Press, Cambridge, Mass. [u.a.], S. XXXVII, 680

Natural Language Processing Group SU. (2008) GATE, A General Architecture for Text Engineering. http://gate.ac.uk (13.03.2008)

Ngai G, Yarowsky D (2000) Rule Writing or Annotation: Cost-efficient Resource Usage for Base Noun Phrase Chunking. Proceedings of the 38th Annual Meeting on Association for Computational Linguistics S. 117 – 125

Ohrn A, Ohno-Machado L (1999) Using Boolean reasoning to anonymize databases. Artif Intell Med 15 (3): S. 235-54

Pommerening K (1995) Pseudonyme - ein Kompromiß zwischen Anonymisierung und Personenbezug. In: Trampisch HJ, Lange S (Hg). Medizinische Forschung - Ärztliches Handeln, MMV Medizin Verlag, München. S. 329-333

Pommerening K, Reng M, Debold P, Semler S (2005) Pseudonymisierung in der medizinischen Forschung - das generische Pseudonymisierung in der medizinischen Forschung - das generische TMF-Datenschutzkonzept. Epidemiol 1 (3): S. Doc17

Rechenberg P, Pomberger G (2006) Informatik-Handbuch. 4., aktualisierte und erw. Aufl, Hanser, München [u.a.], S. 1251

Rijbergen CJ. (1979) Information Retrieval. http://www.dcs.gla.ac.uk/Keith/Preface.html (13.03.2008)

Ruch P, Wagner J, Bouillon P, Baud RH, Rassinoux AM, Scherrer JR (1999) MEDTAG: tag-like semantics for medical document indexing. Proc AMIA Symp S. 137-41

Ruch P, Baud RH, Rassinoux A, Bouillon P, Robert G (2000) Medical Document Anonymization with a Semantic Lexicon. Proc AMIA Symp S. 729-33

Ruch P, Baud R, Geissbühler A (2003) Using Lexical Disambiguation and Named-Entity-Recognition to Improve Spelling Correction in the Electronic Patient Record. Artif Intell Med 29 (1-2): S. 169-84

Sibanda T, He T, Szolovits P, Uzuner O (2006b) Syntactically-informed semantic category recognition in discharge summaries. AMIA Annu Symp Proc S. 714-8

Sibanda TC. (2006a) Was the Patient Cured? Understanding Semantic Categories and Their Relationships in Patient Records. http://groups.csail.mit.edu/medg/ftp/tawanda/THESIS.pdf (13.03.2008)

Sweeney L (1996) Replacing personally-identifying information in medical records, the Scrub system. Proc AMIA Annu Fall Symp S. 333-7

Sweeney L (2002) k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10 (5): S. 557-570

Szarvas G, Farkas R, Busa-Fekete R. (2007) State-of-the-art anonymisation of medical records using an iterative machine learning framework. http://www.jamia.org/cgi/reprint/M2441v1.pdf (13.03.2008)

Tinnefeld M, Ehmann E, Gerling RW (2005) Einführung in das Datenschutzrecht. 4., völlig neu bearb. und erw. Aufl, Oldenburg, München, Wien, S. XX, 770

Tomanek K, Wermter J, Hahn U (2007a) Sentence and Token Splitting Based On Conditional Random Fields. Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics S. 49–57

Tomanek K, Wermter J, Hahn U (2007b) An Approach to Text Corpus Construction which Cuts Annotation Costs and Maintains Reusability of Annotated Data. Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and ComputationalNatural Language Learning S. 486–495

Uzuner O, Luo Y, Szolovits P (2007) Evaluating the State-of-the-Art in Automatic De-identification. J Am Med Inform Assoc 14 S. 550-563

Wallach HM. (2004) Conditional Random Fields: An Introduction. http://www.inference.phy.cam.ac.uk/hmw26/papers/crf_intro.ps (13.03.2008)

Wellner B, Huyck M, Mardis S, Aberdeen J, Morgan A, Peshkin L, Yeh A, Hitzeman J, Hirschman L (2007) Rapidly Retargetable Approaches to De-identification in Medical Records. J Am Med Inform Assoc 14 S. 564-573

Wermter J, Hahn U (2004) Ein annotiertes deutschsprachiges medizinisches Textkorpus. GMDS S. 235-7

Wermter J, Tomanek K, Balzer F (2006) Automatische Erkennung und effiziente Annotation von anonymisierungsrelevanten Begriffen in klinischen Freitexten. GMDS 2006

World Health Organization. (1990) International Classification of Diseases (ICD). http://www.who.int/classifications/icd/en/ (13.03.2008)

7.2 Abbildungsverzeichnis

7.3 Annotationsrichtlinien

Verwendete Tags:

- 1 person
  - 1.1 patient
  - 1.2 medical
- 2 address
  - 2.1 street
  - 2.2 city
  - 2.3 phone
- 3 organisation
  - 3.1 hospital
  - 3.2 dept
  - 3.3 office
- 4 timex
  - 4.1 date
  - 4.2 time
- 5 numex
  - 5.1 length
  - 5.2 vol
  - 5.3 conc
  - 5.4 dose
  - 5.5 percent
  - 5.6 mass
  - 5.7 freq
  - 5.8 other
- 6 device
- 7 implant
- 8 code
  - 8.1 icd
  - 8.2 dsm
  - 8.3 tnm
  - 8.4 ops
  - 8.5 other
- 9 drug
  - 9.1 group
  - 9.2 chemicals
  - 9.3 branddrug

Allgemeine Erläuterungen:

- Das Taggen erfolgt mithilfe der Anwendung MMax2.
- Verschachtelungen sind möglich.
  Beispiel: *Universitätsklinikum **Freiburg***. Hier sind beide Wörter im Tag **hospital** enthalten, die Stadt Freiburg noch mal extra im Tag **city**
- Von Medikamenten abgeleitete Wortschöpfungen, wie Macumarisierung, werden nicht annotiert.

Erläuterungen zu den einzelnen Tags:

1 person

Hierzu gehören die Namen von Patienten sowie Personen, die den Patienten behandeln (medical professionals). Inhalt des Tags sind Vorname mit Name sowie Anrede und Titel. Berufsbezeichnungen (z.B. „der behandelnde Arzt“) gehören nicht dazu.

1.1 patient

Dieses Tag dient der Annotation von Patientennamen.

Beispiele:

- Begleitend wurde *Frau Romann* über die Naturheilkunde-Ambulanz unseres Hauses mitbetreut .
- *Frau Dr. Mielke* konnte in deutlich gebesserten Allgemeinzustand am 05.02.2002 entlassen werden.

1.2 medical

Mit dem Tag **medical** werden Namen von so genannten medical professionals gekennzeichnet, also Ärzten, Schwestern, etc.

Beispiele:

- Histo (*Doktor Quasem* , Biberach) : mässig diff. Adeno-Carcinom
- Der diensthabende Arzt, *Dr. Lohse*, führt die notfallmäßige Intubation um 16:55h durch.

2 address

Tags dieser Gruppe werden sowohl für Adressen von Patienten und Institutionen als auch zur Annotation von anderen Orten (z. B. Geburtsorten) genutzt.

2.1 street

Auch Hausnummern werden in das street-Tag eingeschlossen.

Beispiel:

- Laurenz Prächtel geb. 10.02.1944,
  Joschka-Fischer-Strasse 2b, 90764 Neustadt

2.2 city

Geht der Stadt eine Postleitzahl voran, gehört auch diese zum Tag.

Beispiel:

- Laurenz Prächtel geb. 10.02.1944,
  Joschka-Fischer-Strasse 2b, *90764 Neustadt*

2.3 phone

Zu diesem Tag gehören auch Faxnummern und ähnliches

Beispiel:

- Wir bitten um telefonische Terminvereinbarung unter *0121-3849240*.

3 organisation

Diese Kategorie umfasst Behandlungsorte.

3.1 hospital

Mit dem Tag **hospital** werden Namen von Krankenhäuser, Kurkliniken, etc. gekennzeichnet. Abteilungen, wie z.B. „Frauenklinik“ werden nicht als Krankenhaus-, sondern als Stationsname (vgl. Tag **dept**) gehandhabt.

Beispiel:

- *Universitätsklinikum Neustadt an der Elbe*
  Günzelstraße 6
  28493 Neustadt an der Elbe

Hinweis: Im obigen Beispiel ist zu beachten, dass der Ort Neustadt an der Elbe zusätzlich noch als **city** annotiert werden muss.

3.2 dept

Gemeint sind Stationen und Ambulanzen eines Krankenhauses

Beispiel:

- Wir bitten um eine ambulante WV in unserer *HCC-Ambulanz* am 25.08.02 um 8:00 Uhr zur Durchführubng eines CT´s zur Verlaufskontrolle.
- Wir überweisen Herrn Winckelmann an die *Universitäts-Hautklinik* zur Abklärung einer Effloreszens.
- Frau Schuhmann wurde am 17.8. auf die *Station Scheuermann* verlegt.

3.3 office

Das **office**-Tag dient der Markierung von Arztpraxen

Beispiel:

- Die ambulante Weiterbehandlung erfolgt wie gehabt durch die *Praxis Dr. Meisenbach*

4 timex

Bei der Annotation von Zeit- und Datumsangaben wird zusätzlich festgelegt, ob es sich um Intervalle handelt (Zusatzoption **period**) und auch, ob die Angaben absolut oder relativ zu verstehen ist. Bei solchen relativen anonymisierungsrelevanten Textbestandteile (z.B. „in drei Wochen“) lässt sich der Zeitpunkt nur aus dem Kontext erschließen. Anonymisierungsrelevant sind lediglich absolute Angaben.

4.1 date

Datumsangaben können in verschiedenster Form vorkommen. In das Tag wird neben einleitenden Adverbien (z.B. seit, bis) auch die Abkürzung ED (Erstdiagnose) aufgenommen.

Beispiele für absolute Datumsangaben:

- *ED 08'02*
- *ED 07/03*
- *seit' 96*
- *ab 2004*
- *im Winter 2003*
- *2.3.1989*
- *Donnerstag, der 02.1.2004*
- *02.03'93*
- *vom 2.3. bis 3.4.01* (Zusatzoption **period**)
- *4.5.-6.7.05*
- 1. Zyklus: d1 am *08.07.02*, d8 am *16.07.02*, d15 am *23.07.02*
- *Im Juni*

Beispiele für relative Datumsangaben:

- *seit insgesamt zwei Wochen*
- *von zwei Wochen*
- *gestern*
- *vor Jahrzehnten*
- *bis 2 Tage vor der OP*
- *letzte Woche*
- *heute Mittag*
- *in einem Jahr*
- *Tag 4 der Chemotherapie*
- *d1-d5* (Zusatzoption **range**)
- der *61-jährige* Patient
- die *tägliche* Gabe von ASS
- der Patient wurde *70 Jahre* alt
- Therapiestand: Chemotherapie mit Gemzitabin 1000mg/m2 (*Tag 1,8,15*)
- 1. Zyklus: *d1* am 08.07.02, *d8* am 16.07.02, *d15* am 23.07.02

4.2 time

Dieses Tag wird analog zum Tag **date** gebraucht, jedoch nur dann, wenn die Zeitangabe weniger als 24 Stunden beträgt. Man beachte die unterschiedlichen Arten, in den Uhrzeiten angegeben werden.

Beispiele für absolute Zeitangaben:

- Um *17.05h* rief die Ehefrau den Notarzt
- Nach der zweistündigen Operation Auftreten einer Angina pectoris-Symptomatik um *18:55*
- Der Patient wurde um *viertel nach sieben* auf die Intensivstation verlegt

Beispiele für relative Zeitangaben:

- Nach der *zweistündigen* Operation Auftreten einer Angina pectoris-Symptomatik um 18:55
- Etwa *fünf Minuten vor* Auftreten des anaphylaktischen Schocks beklagte der Patient ...
- *3 min später* trat eine vorläufige Besserung ein.
- *bis 3 min*

5 numex

Enthält numerische Werte mit den dazugehörigen Maßeinheiten. Generell wird kein Unterschied gemacht, ob Zahlen als Ziffern oder Worten angegeben werden (z.B. „*drei bis vier mal pro Tag*“ = „*3-4/d*“). Wie bei der Kategorie **timex** kann bei anonymisierungsrelevanten Textbestandteilen festgelegt werden, ob diese relativ oder absolut sind bzw. ob ein Intervall vorliegt.

5.1 length

Längen

Beispiel:

- Ein bis *2 cm* im Durchmesser großes Resektat wird lamelliert und komplett eingebettet.
- Der Ösophagus wurde um *einen cm* verkürzt. (Zusatzoption **relative**)

5.2 vol

Volumen

Beispiel:

- Der Patient verlor während der etwa einstündigen Operation *1,5-2 l* Blut. (Zusatzoption **period**)

5.3 conc

Konzentrationen

Beispiele:

- Leukozyten *7,4 Tsd/µl*
- Erythrozyten *3,91 Mio/µl*
- Hämoglobin *11,4 g/dl*
- *1000mg / m^2*

5.4 dose

Dosis eines Medikamentes

Beispiele:

- Sap Simplex *40 Tropfen*
- Pantozol 20 *1-0-0*
- *eine ganze Tablette*
- *1x täglich*
- *20-20-20Trpf.*

5.5 percent

Prozent

Beispiel:

- Hämatokrit *34,7 %*
- *80-85%* (Zusatzoption **period**)

5.6 mass

Masse

Beispiel:

- Das Gewicht habe bis auf *87 kg* zugenommen.
- Gewichtszunahme *von 44,1 auf 44,6 kg* (Zusatzoption **period**)

5.7 freq

Frequenzen

Beispiel:

- Puls *68/min*
- Stuhlfrequenz von *14 pro Tag*

5.8 other

Andere Einheiten

Beispiele:

- Radiatio (stereotaktisch) 8-9/00 mit einer Gesamtdosis von *39 Gy*
- Brandwunde von einer Fläche von *3 x 3 cm* ; *2x2x3cm*; *3cm^2*

6 device

Medizinische Geräte, z. B. CT-Apparate, OP-Geräte, etc.

Beispiel:

- Spiral-CT vom 2.3.04 mit *RÖNTGOMAT2000*
- *Diathermiemesser*
- *Dymont – Bronchoskop*
- *Redon Drainage*

7 implant

Implantate, Prothesen

Beispiele:

- Z.n. endoskopischer Papillotomie und Einlage einer transpapillären Gallengangsendoprothese (11,5 F/12 cm, Typ *Tannenbaum*) 08/00
- Z.n. Mitralklappenersatz (Bioprothese-*Carpentier/Edwards*)

8 code

Codes zur Diagnoseverschlüsselung, Tumorklassifikationen, etc. Der Name des Codes (z.B. ICD) gehört nicht zum Tag.

8.1 icd

International Classification of Diseases

Beispiele:

- Postoperative benigne Choledochusstenose (ICD *JX-K71.0*)
- Mitralinsuffizienz I. Grades (*JX-I34.0*)

8.2 dsm

Diagnostic and Statistical Manual of Mental Disorders

Beispiele:

- Z. n. depressiver Störung (*296.26*)

8.3 tmn

Tumorklassifikationssystem (Tumor-Metastasis-Node invasion System)

Beispiele:

- Schleimbildendes Adenokarzinom des Pankreaskopfes ED 4/99, (IX-C25.0), initial *T3 Nx M1* (multiple Lebermetastasen)
- Cholangiozelluläres Karzinom: peripher, Stadium *cT4cN1Mx*
- *T4N1M1* (ED 12.10.01)

8.4 ops

OP-Codes

Beispiele:

- Implantation einer Totalprothese (*5-820.01*)

8.5 other

Andere Klassifikationen

9 drug

Kategorie zur Annotation von Medikamenten. Als Wirkstoffe und Handelsnamen sollen nur Wörter in Frage kommen, die keine Pluralbildung zu lassen.

Beispiele:

annotationsrelevant (keine Pluralbildung möglich)	nicht annotationsrelevant (Pluralbildung möglich)
Lebertran Valium Diazepam Insulin (Insuline bedeutet Insulinarten!) Heparin	Lebertrankapsel Vitamintablette Heparinspritze Insulindosis

Außerdem sollen keinen Neologismen annotiert werden, wie z.B. „der macumarisierte Patient“.

9.1 group

Tag für Substanzklassen und Gattungsbezeichnungen

Beispiele:

- Absetzen des *Betablockers* bei orthostatischem Schwindel.
- Arterielle Hypertonie, Unverträglichkeit (unsystematischer Schwindel) auf *ACE-Hemmer*
- *Lebertrankapseln*
- *Kopfschmerztabletten*
- Die Gabe von *Vitaminen* wird empfohlen

9.2 chemical

Dieses Tag steht für Medikamentenwirkstoffe und chemische Bezeichnungen. Existiert eine Bezeichnung gleichzeitig als Wirkstoff und Handelsname, wird sie im Zweifelsfall als **chemical** getaggt.

Beispiele:

- *Metamizol-Natrium* (z.B. Novalgin) Trp. 20-20-20-20
- Der Patient entwickelte eine Unverträglichkeit auf *Acetylsalicylsäure*.
- Wir empfehlen die tägliche Gabe von 1g *Vitamin C*
- *Heparin-Natrium* Braun 25 000 I.E./5 ml
- *Leinsamen*

9.3 branddrug

Mit diesem Tag werden Handelsnamen von Medikamenten annotiert

Beispiele:

- Metamizol-Natrium (z.B. *Novalgin*) Trp. 20-20-20-20
- Medikation: *ASS* 300 0-1-0, *Zyloric* 300 0-0-1

7.4 Bundesdatenschutzgesetz (Auszug)

BDSG § 3 Weitere Begriffsbestimmungen

(1) Personenbezogene Daten sind Einzelangaben über persönliche oder sachliche Verhältnisse einer bestimmten oder bestimmbaren natürlichen Person (Betroffener).
(2) Automatisierte Verarbeitung ist die Erhebung, Verarbeitung oder Nutzung personenbezogener Daten unter Einsatz von Datenverarbeitungsanlagen. Eine nicht automatisierte Datei ist jede nicht automatisierte Sammlung personenbezogener Daten, die gleichartig aufgebaut ist und nach bestimmten Merkmalen zugänglich ist und ausgewertet werden kann.
(3) Erheben ist das Beschaffen von Daten über den Betroffenen.
(4) Verarbeiten ist das Speichern, Verändern, Übermitteln, Sperren und Löschen personenbezogener Daten. Im Einzelnen ist, ungeachtet der dabei angewendeten Verfahren:
1. Speichern das Erfassen, Aufnehmen oder Aufbewahren personenbezogener Daten auf einem Datenträger zum Zwecke ihrer weiteren Verarbeitung oder Nutzung,
2. Verändern das inhaltliche Umgestalten gespeicherter personenbezogener Daten,
3. Übermitteln das Bekanntgeben gespeicherter oder durch Datenverarbeitung gewonnener personenbezogener Daten an einen Dritten in der Weise, dass
a) die Daten an den Dritten weitergegeben werden oder
b) der Dritte zur Einsicht oder zum Abruf bereitgehaltene Daten einsieht oder abruft,
4. Sperren das Kennzeichnen gespeicherter personenbezogener Daten, um ihre weitere Verarbeitung oder Nutzung einzuschränken,
5. Löschen das Unkenntlichmachen gespeicherter personenbezogener Daten.

(5) Nutzen ist jede Verwendung personenbezogener Daten, soweit es sich nicht um Verarbeitung handelt.
(6) Anonymisieren ist das Verändern personenbezogener Daten derart, dass die Einzelangaben über persönliche oder sachliche Verhältnisse nicht mehr oder nur mit einem unverhältnismäßig großen Aufwand an Zeit, Kosten und Arbeitskraft einer bestimmten oder bestimmbaren natürlichen Person zugeordnet werden können.
(6a) Pseudonymisieren ist das Ersetzen des Namens und anderer Identifikationsmerkmale durch ein Kennzeichen zu dem Zweck, die Bestimmung des Betroffenen auszuschließen oder wesentlich zu erschweren.
(7) Verantwortliche Stelle ist jede Person oder Stelle, die personenbezogene Daten für sich selbst erhebt, verarbeitet oder nutzt oder dies durch andere im Auftrag vornehmen lässt.
(8) Empfänger ist jede Person oder Stelle, die Daten erhält. Dritter ist jede Person oder Stelle außerhalb der verantwortlichen Stelle. Dritte sind nicht der Betroffene sowie Personen und Stellen, die im Inland, in einem anderen Mitgliedstaat der Europäischen Union oder in einem anderen Vertragsstaat des Abkommens über den Europäischen Wirtschaftsraum personenbezogene Daten im Auftrag erheben, verarbeiten oder nutzen.
(9) Besondere Arten personenbezogener Daten sind Angaben über die rassische und ethnische Herkunft, politische Meinungen, religiöse oder philosophische Überzeugungen, Gewerkschaftszugehörigkeit, Gesundheit oder Sexualleben.
(10) Mobile personenbezogene Speicher- und Verarbeitungsmedien sind Datenträger,
1. die an den Betroffenen ausgegeben werden,
2. auf denen personenbezogene Daten über die Speicherung hinaus durch die ausgebende oder eine andere Stelle automatisiert verarbeitet werden können und
3. bei denen der Betroffene diese Verarbeitung nur durch den Gebrauch des Mediums beeinflussen kann.

7.5 Lebenslauf

NAME	Felix Balzer
GEBOREN	11.3.1981 in Hamburg
STAATSANGEHÖRIGKEIT	deutsch
ANSCHRIFT	Lehener Straße 11, 79106 Freiburg
SCHULBILDUNG	– Grundschule Karlshöhe, Hamburg 1987 – 1991 – Peter-Petersen-Gesamtschule, Hamburg 1991 – 1997 – St. Xavier High School, Cincinnati, Ohio, USA 1997 – 1998 – Albert-Schweitzer-Gymnasium, Hamburg 1998 – 2000
ZIVILDIENST (ANDERER DIENST IM AUSLAND)	– Association Le Champ de la Croix Institut médico-éducatif « Les Allagouttes » 68370 Orbey, Frankreich September 2000 – September 2001
STUDIUM DER HUMANMEDIZIN (STAATSEXAMEN)	– Universität Rostock Oktober 2001 – März 2003 – Albert-Ludwigs-Universität Freiburg April 2003 – März 2008 – Université de Paris Sud XI (ERASMUS-Stipendium) Oktober 2004 – September 2005
PRAKTISCHES JAHR	– Anästhesiologie an der HELIOS-Klinik Titisee-Neustadt Albert-Ludwigs-Universität Freiburg Februar – Mai 2007 – Innere Medizin am Hôpital Cochin Paris Université Paris Descartes, Frankreich Mai – September 2007 – Chirurgie am New York Downtown Hospital Weill Medical College of Cornell University, USA September – November 2007 – Chirurgie am Centre Hospitalier Universitaire de Guadeloupe Université des Antilles et de la Guyane, Frankreich November 2007 – Januar 2008
STUDIUM DER INFORMATIK (MASTER OF COMPUTER SCIENCE)	– Fernuniversität in Hagen seit Oktober 2005

Danksagung

Der größte Dank gebührt meinem Doktorvater und Betreuer Prof. Dr. med. Stefan Schulz. Ich danke ihm für die freundliche Aufnahme in die Forschungsgruppe, die Überlassung und Betreuung dieser Arbeit und für das in mich gesetzte Vertrauen und die bereitwillige Förderung. Ich habe es als sehr konstruktiv empfunden, dass jede auftretende Frage umgehend und mit viel Zeit beantwortet wurde.

Herrn Joachim Wermter und Frau Katrin Tomanek (Jena University Language & Information Engineering Lab) danke ich außerordentlich für ihre Hilfe bei der Einarbeitung in die Experimente und Durchführung der vorliegenden Studien sowie ihre tatkräftige Unterstützung bei technischen und wissenschaftlichen Fragen.

Frau Prof. Dr. rer. nat. Katharina Nübler-Jung (Chirurgische Universitätsklinik Freiburg) danke ich herzlich für die freundliche und unkomplizierte Übernahme des Zweitgutachtens.

Ein besonderer Dank gilt meiner Mutter für ihre liebevolle Unterstützung.