gmds · LEIPZIG 2006

Jena University Language & Information Engineering Lab

Automatische Erkennung und effiziente Annotation von anonymisierungsrelevanten Begriffen in klinischen Freitexten

Joachim Wermter **Katrin Tomanek** **Felix Balzer**

StemNet

SEMANTIC MINING

Boot Strep

Einleitung

- Vertrauliche Behandlung aller sensiblen Daten eines Patienten
- Großes Interesse an klinischen Patienten-Daten
  - Fallbeschreibungen aus der Praxis
  - großer Fundus an Daten erlaubt gezielte Fragen
- Aber: Vertraulichkeit schwer zu wahren
  - Manuelle Anonymisierung sehr aufwändig (vor allem bei großen Datenmengen)
  - Durch Fehler trotzdem Rückschlüsse auf Identität des Patienten möglich (Sweeney [1996]: 6 %)

Fragestellung

- Ziel: automatisierte Anonymisierung medizinischer Daten
  - Vorverarbeitungsschritt
  - macht manuelle Anonymisierung effizienter
  - De-identification challenge (AMIA 2006)
- Gängige anonymisierungsrelevante Begriffe:
  - Patientenname, -adresse, Arzt, Klinik, Abteilung, etc.
  - Viele befinden sich in strukturierten Abschnitten klinischer Dokumente (Dateikopf, etc.) oder in den Stammdaten
  - Input für eine abgleichungsbasierte Anonymisierung im viel unzugänglicheren unstrukturierten Freitext-Teil

STÄDTISCHE KLINIKEN NEUSTADT AM MAIN
Abt. Innere Medizin II • Holzhauser Str. 55 • 90766 Neustadt

Herrn
Dr. med. Martin Müller
Willy-Brandt-Str. 26

90764 Neustadt

MEDIZINISCHE KLINIK UND POLIKLINIK
Abteilung Innere Medizin 6
Hepatologie und Endokrinologie
Ärztlicher Direktor Prof. Dr. E. Baum

	Telefon / Fax
Endoskopie	0182-2170-3303 / 3259
Sonographie	3346 / 3259
Crohn- / Colitis-Ambulanz	3308 / 3447
Leber- und Transplant-Ambulanz	3308 / 3447
Magen-Darm-Ambulanz	3308 / 3447
TIPS-Ambulanz	3261 / 3259
Diabetologische Ambulanz	3512 / 3656
Endokrinologische Ambulanz	3508 / 3656
HIV (stationär)	3463 / 3287
Infektiologische Ambulanz	3308 / 3447

Neustadt 08.08.1997 hei/bi

Clarissa Caesar, geb. 01.12.1959,
Friedrichstrasse 5, 90764 Neustadt

Sehr geehrter Herr Kollege,

wir berichten Ihnen nachfolgend über o.g. Patientin, die sich am 03.07.1997 in unserer Ambulanz vorstellte.

Fragestellung

- Viel schwieriger ist die Identifizierung von Daten / Begriffen, die sich nicht aus strukturierten Dateiköpfen oder Stammdaten ableiten lassen
  - Befinden sich hauptsächlich im unstrukturierten Freitext-Teil eines klinischen Dokuments
  - Typischerweise durch große lexikalische Vielfalt gekennzeichnet
  - Manuelle Auflistung aller Muster inhärent unvollständig
- Datums- und Zeitangaben: erlauben durchaus Rückschlüsse auf Identität eines Patienten

Sehr geehrter Herr Kollege,

wir berichten Ihnen nachfolgend über o.g. Patientin, die sich am 03.07.1997 in unserer Ambulanz vorstellte.

Diagnosen: Hepatitis-B-Virusinfektion

Die **Anamnese** der Patientin dürfen wir als bekannt voraussetzen. Die ED der Hepatitis B war am 28.03.96. Sie war das erste Mal in der Leberambulanz am 25.04.96. Damals zeigte sich HBsAg positiv, Anti-HBc-IgG positiv, HBeAg negativ und Anti-HBeAk positiv. Die Transaminasen waren normal. Es konnte die Diagnose eines Carrierstatus gestellt werden. In einer Kontrolle bei Ihnen im Februar 97 zeigte sich die selbe serologische Konstellation. Die Transaminasen waren auch normal, allerdings hatten sie einen positiven Hepatitis-B-Virus-DNS-Nachweis. Nach weiteren sechs Monaten stellte sich die Patientin am 03.07.97 erneut in der Leberambulanz vor.
Systematische Anamnese: Die Patientin fühlt sich häufig schlapp und müde. Ansonsten hat sie keine Beschwerden. Insbesondere sind im letzten Jahr keine neuen Symptome aufgetaucht.

Status:
168 cm groß, 69 kg schwer, RR 140/90 mmHg, Puls 66/min., regelmäßig. 2/6 Holosystolikum. Hepatojugulärer Reflux negativ, sämtliche peripheren Pulse gut. Lungen frei. Abdomen unauffällig. Leber normal. Keine Leberhautzeichen. Milz nicht palpabel. Kein Ikterus.

Labor:
Eine Kopie der Laborwerte liegt bei. Zusammenfassend zeigt sich ein normales Blutbild, eine normale Gerinnung sowie normale Leberwerte. Das AFP ist ebenfalls normal.
In der Serologie zeigt sich nun HBsAg positiv, HBsIgG grenzwertig, Anti-HBc-IgG positiv, HBeAg negativ, Anti-HBe positiv, HBV-DNA negativ.

Sonographie d. Abdomens:

Automatische Begriffserkennung

- (Computerlinguistische) Methoden zur automatischen Begriffserkennung (Entity Recognition)
  - Basierend auf maschinellem Lernverfahren (CRF – Conditional Random Fields)
  - Klassifizierer leiten Muster aus Trainingsdaten ab
  - Wortsequenzen (z.B. Datums-/Zeitangaben) eines Satzes werden mit gelernten Markern versehen
- CRFs sind äußerst robust (z.B. bei Erkennung von lexik. Variationen) mit hohen Erkennungsraten
  - bei genügend qualitativ hochwertigen Trainingsdaten!

Automatische Begriffserkennung

- Problem: Erstellung (= Annotation) solcher Trainingsdaten sehr teuer (zeit-/arbeitsintensiv)
  - Daher: oft schwierig, ausreichend annotiertes Trainings-material bereit zu stellen
  - Zudem: selbst bei größeren Textmengen oft nur geringe Dichte an relevanten Begriffen (= positive Lernbeispiele)
  - Folglich: menschliche Annotatoren müss(t)en große Textmengen sichten und annotieren

Lösung: Active Learning (AL)

- Intelligente Selektionsstrategie
  - Gezielte, iterative Bereitstellung der informativsten Textdaten zur Annotation
- Verfahren in der jeder AL-Runde
  - Trainieren eines Komitees von Klassifizierern auf unterschiedlichen Teilbereichen schon annotierter Textdaten
  - Unterschiedlich trainierte Klassifizierer identifizieren in noch nicht annotierten Textdaten die zu erkennenden Begriffe
  - Vergleichen der von jedem Klassifizierer vorhergesagten Begriffe (hier: Zeit-/Datumsangaben) auf Satzebene
  - Selektion der Sätze mit höchster Nicht-Übereinstimmung zur nachfolgenden manuellen Annotation
    - Selektion besonders informativer Trainings-Beispiele
    - Keine Annotierung von uninformativen Sätzen
  - Beenden des AL-Prozesses bei geringer Nicht-Übereinstimmung

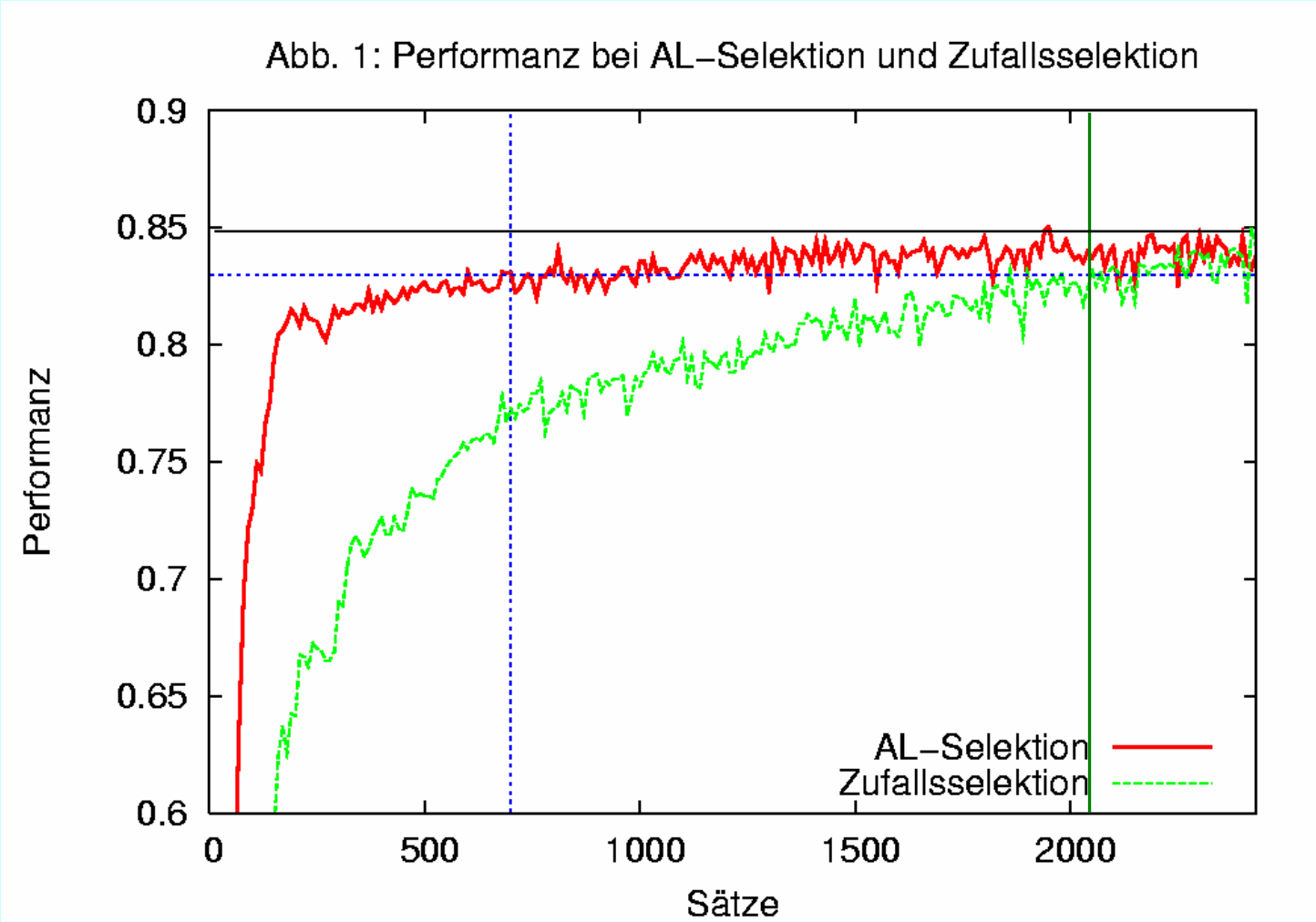
Experimentelle Daten

- FRAMED-Korpus [Wermter und Hahn 2004]
  - Heterogene Textmenge klinischer Dokumente (Arztbriefe, Pathologie- / Histologie- / OP-Berichte)
  - Insgesamt 3.486 Sätze mit 50.655 Wörtern
  - Annotation aller vorkommenden Datums- und Zeitangaben durch Medizinstudenten nach vorgegebenen Richtlinien
- Korpus-Split / -Aufteilung im Verhältnis 70:30
  - AL-Simulationskorpus (2.440 Sätze)
  - Goldstandard (1.046 Sätze)

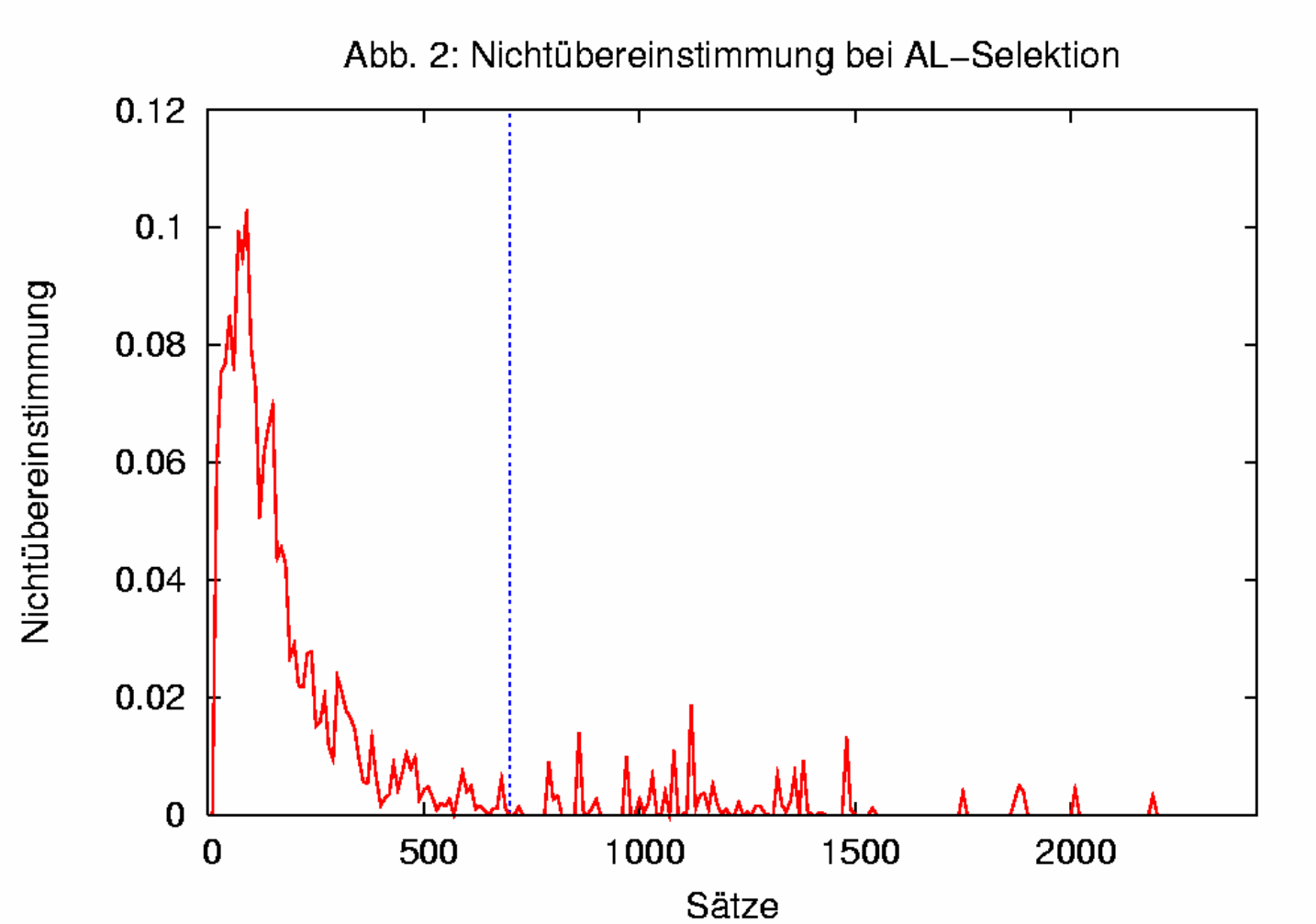
Ablauf AL-Simulationsexperiment

- Selektion in jeder AL-Simulationsrunde:
  - Trainieren eines Komitees aus 3 CRF-Klassifizierern auf unterschiedl. Teilbereichen des schon annotierten Teil des Simulationskorpus
    - Jeder Klassifizierer wird auf 2/3 der schon annotierten Daten trainiert
  - Vorhersage auf nicht annotiertem Teil
  - Bereitstellung der 10 Sätze mit der höchsten Nicht-Übereinstimmung zur weiteren simulierten manuellen Annotation
- Zufallselektion (Baseline) in jeder Runde:
  - Die Sätze zur weiteren Annotation zufällig ausgewählt
- Performanzbestimmung (F1-Score) auf Goldstandard nach jeder Runde
  - Trainieren eines Klassifizierers auf bisher annotiertem Simulationskorpus
  - Insgesamt: Fünf Simulationsläufe mit Mittelung der Performanz

Resultate: Performanz AL-Selektion und Zufallsselektion



Resultate: Nicht-Übereinstimmung bei AL-Selektion



Diskussion und Schlussfolgerungen

- Machbarkeitsstudie: Automatische medizinische Begriffserkennung als Vorbearbeitungsschritt zur Anonymisierung von klinischen Patientendaten
  - Schwer zu erfassende (weil heterogene) Datums- und Zeitangaben in klinischen Freitexten
- AL-Selektionsstrategie stellt effizient Trainingsdaten für maschinelles Lernverfahren bereit
  - Annotation von weniger als 1/3 der vorhandenen Textmenge bei fast gleicher Performanz
- Übertragung auf andere medizinische Begriffe von Interesse in klinischen Freitexten
  - Maßangaben, Diagnosen, Medikamenten-/Substanznamen, etc.
  - Automatische Informationsextraktion / Semantische Suche

gmds · LEIPZIG 2006

Jena University Language & Information Engineering Lab

Automatische Erkennung und effiziente Annotation von anonymisierungsrelevanten Begriffen in klinischen Freitexten

Joachim Wermter **Katrin Tomanek** **Felix Balzer**

StemNet

SEMANTIC MINING

Boot Strep