TOP-DOWN INFORMATIONSEXTRAKTION AUS KLINISCHEN TEXTEN FÜR DIE SEKUNDÄRNUTZUNG DER ELEKTRONISCHEN PATIENTENAKTE

Kreuzthaler M[1], Schulz S[1], Berghold A[1]

Kurzfassung

Das Erschließen eingebetteter Information in teilstrukturierten Daten in Kombination mit nichtstandardisiertem Text und Qualitätsaspekten in der medizinischen Routinedokumentation stellt spezielle Herausforderungen an Methoden der Informationsextraktion. Anhand eines prototypischen Informationsextraktionssystems, basierend auf UIMA und regulären Ausdrücken wird untersucht, in wieweit sich zuvor definierte Merkmale aus dem medizinischen Freitext extrahieren lassen.

Abstract

Accessing clinical information embedded in semi-structured data in combination with non-standardized text, together with quality aspects in routine documents provides special methodological challenges for information extraction. Building a prototypical information extraction system based on UIMA and regular expressions it is examined to what extent predefined attributes can be extracted from medical free text.

Keywords – Information Extraction, Clinical Narrative, Secondary Use, UIMA

1. Einleitung und Motivation

Teilstrukturierte Daten und nichtstandardisierter Text sind in klinischen Informationssystemen (KIS) nach wie vor weit verbreitet. Das Erschließen der eingebetteten Information in diesen Datenstrukturen in Kombination mit Qualitätsaspekten der medizinischen Routinedokumentation stellt spezielle Herausforderungen an Methoden der Informationsextraktion. Zu berücksichtigen sind Eigenheiten des Medizinjargons wie Synonymie, Homonymie, Abkürzungen und Akronyme, Komposita, Schreibvarianten, Tipp- und Rechtschreibfehler, Sprachregister, Telegramm-Stil, Mehrsprachigkeit (Englisch, Deutsch, Latein) und nichtstandardisierte numerische Ausdrücke. Die inhaltliche Erfassung dieser Textbestände erlangt mehr und mehr an Bedeutung. Vor allem das Zusammenführen von Datenbeständen wie z.B. Biobankdaten in Kombination mit klinischen Daten wird als vielversprechende Quelle für das Erschließen neuen medizinischen Wissens angesehen[2]. Auch für Studien, wo Patientenkollektive mit gewissen Expositionen gesucht werden, stellt die

[1] Institut für Medizinische Informatik, Statistik und Dokumentation, Medizinische Universität Graz, Österreich

[2] http://emerge.mc.vanderbilt.edu/

Elektronische PatientInnen-Akte (EPA) eine unverzichtbare Quelle dar[1]. Die Aussagekraft der Studien hängt dabei wesentlich von der Qualität der medizinischen Routinedokumentation ab. Für das Verarbeiten dieser teil- bzw. unstrukturierten textuellen Datenbestände gibt es verschiedene Ansätze aus dem Bereich des *Natural Language Processing* (NLP) wobei diese Herausforderungen auch im akademischen Umfeld wahrgenommen werden und mit Challenges wie z.B. dem TREC Medical Records Track[2] oder i2b2[3] angesprochen werden. Dabei etabliert sich Apache UIMA[4] (*Unstructured Information Management Architecture*) immer mehr als Standard. An medizinische Sprache angepasst sind ferner Systeme wie cTakes[5] und MedLEE[6]. Bei der Verarbeitung von „Big Data" können die Apache-Projekte Hadoop[7] oder Mahout[8] in Betracht gezogen werden. Es werden aber auch No-SQL Ansätze und Semantic Web-Technologien verstärkt berücksichtigt [1]. Anwendungen dieser Technologien auf Freitexte werden unter anderem in [2-4] diskutiert.

2. Zielsetzung

Im Rahmen dieser Arbeit wird prototypisch versucht, Daten einer klinischen Studie, die bereits in einem Wissenschafts-Dokumentationssystem erfasst wurden, maschinell unter Einsatz ausdrucksstarker regulärer Ausdrücke (Top-Down-Ansatz) aus freitextlicher Routinedokumentation zu extrahieren und die Güte dieser Extraktion zu bewerten. Als Framework für die Datenverarbeitung wird UIMA herangezogen. Die Daten, die für die Studie aus der Routinedokumentation benötigt wurden, lagen zu 50% im Freitext vor, die anderen 50% in strukturierter Form.

3. Methoden und Daten

3. 1. UIMA

UIMA ist ein Framework für die Verarbeitung unstrukturierter, insbesondere textueller Daten primär von IBM [5] entwickelt und seit 2006 in die Apache Software Foundation eingegliedert. Die wichtigsten Komponenten der Architektur sind in *Abbildung 2* dargestellt. Kern dabei bildet eine *Collection Processing Engine* bestehend aus einem *Collection Reader*, einer oder mehreren *Analysis Engines* und einem *CAS Consumer*. Der *Collection Reader* dient dem Erfassen der unstrukturierten Datenbestände wie z.B. XML oder pdf Dateien. Die eingelesenen Daten werden anschließend an eine Abfolge von *Analysis Engines* übergeben (*Aggregate Analysis Engine*) wobei eine *Analysis Engine* jeweils einen Verarbeitungsschritt in einer NLP-Pipeline übernehmen kann. Der annotierte Datensatz wird an einen *CAS Consumer* weiter gereicht, welcher typischerweise die erstellten Annotation als Suchmaschinen-Indizes bereitstellt oder in strukturierter Form in eine Datenbank speichert. UIMA im Kontext der Verarbeitung von medizinischen Freitexten wird in [6] beschrieben.

[1] http://www.ehr4cr.eu/
[2] http://trec.nist.gov/pubs/call2011.html
[3] https://www.i2b2.org/NLP/
[4] http://uima.apache.org/
[5] http://ctakes.apache.org/
[6] http://healthfidelity.com/technology
[7] http://hadoop.apache.org/
[8] http://mahout.apache.org/

3. 2. Goldstandard und Datenverarbeitung

Die Extraktion der medizinischen Freitexte wurde für eine Auswahl von StudienpatientInnen in einem *Extract Transform Load* (ETL) Schritt aus dem KIS durchgeführt und in einen Test- und Trainingsdatensatz aufgeteilt. Daten von 39 PatientInnen wurden für die Entwicklung einer Menge an regulären Ausdrücken pro Attribut verwendet. Aus weiteren 39 PatientInnen wurde der Testdatensatz gebildet. Als Goldstandard wurden die bereits manuell erhobenen Attribute aus dem Wissenschafts-Dokumentationssystem exportiert und mit der maschinellen Extraktion verglichen. Ein wichtiger Punkt bei der Evaluierung ist die Tatsache, dass alle vorkommenden Werte eines bestimmten Attributes (z.B. Gewicht) zu einem/einer Patienten/Patientin maschinell extrahiert wurden. Stimmt aus dieser maschinell extrahierten Menge mindestens ein Wert mit dem erfassten Wert im Wissenschafts-Dokumentationssystem (Goldstandard) für diesen Patienten überein, wurde dieser als erkannt bewertet (*best case scenario*). Die verschiedenen Komponenten der Evaluierungspipeline sind in *Abbildung 1* zu sehen.

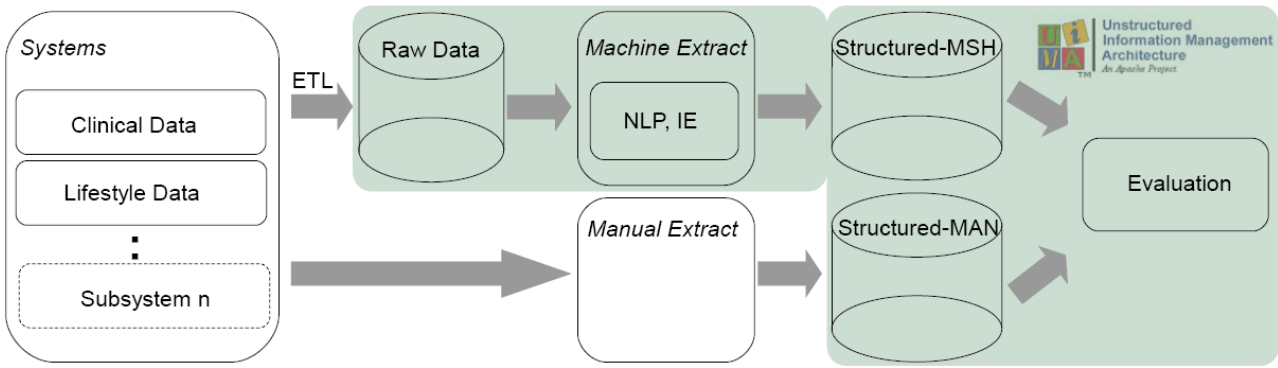

Abbildung 1: Evaluierungspipeline

Die grün hinterlegten Komponenten aus *Abbildung 1* wurden in weiterer Folge auf Strukturen des UIMA Frameworks (*Abbildung 2*, gelbe Elemente) abgebildet:

Source**:** Die extrahierten Dokumente, die in verschiedenen Dokumenttypen vorliegen (z.B. xml, pdf) formen den zu bearbeitenden Dokumentenpool (*Raw Data*). *Raw Data* dient als Input für den Collection Reader.

Collection Reader**:** *Raw Data* wird mit der Java-Klasse `PDFCollectionReader` unter zu Hilfenahme von iText[1] erfasst, wobei alle zu einem/einer Patienten/Patientin gehörenden Befundtexte in einem zu analysierenden Text zusammengefasst werden.

Analysis Engine: Für jedes Attribut existiert eine eigene Analysis Engine. Für manche Attribute wurden zusätzlich noch Value Guards implementiert, die der dementsprechenden Analysis Engine vorgeschaltet und auf das Auffinden spezieller Textabschnitte ausgelegt sind *(Machine Extract).*

CAS Consumer: Eine Java-Klasse `EvaluationWriterCasConsumer` vergleicht die strukturierten maschinell extrahierten Attribute (*Structured-MSH*) mit den manuell extrahierten (*Structured-MAN,* CSV-Export aus dem Wissenschafts-Dokumentationssystem) und evaluiert das Ergebnis des kompletten Extraktionsprozesses.

[1] http://itextpdf.com/

Structured Results: Die strukturierten Daten werden in eine Textdatei geschrieben, wobei zeilenweise die Ergebnisse der maschinellen Extraktion und der manuellen Extraktion untereinander ausgegeben werden.

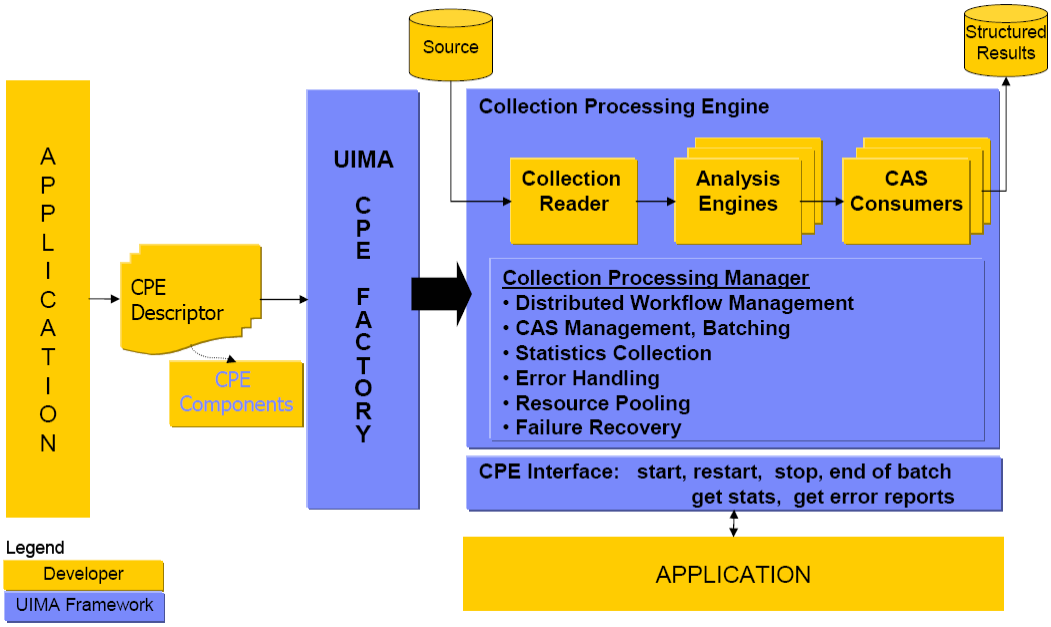

Abbildung 2: UIMA Framework[1]

4. Ergebnisse

Die erzielten Ergebnisse für den Trainings- und Testdatensatz sind in *Tabelle 1* dargestellt, wobei für beide Datensätze Erkennungsraten von über 0,9 erzielt wurden. Von 12 pro Patient erhobenen Attributen (n=468) im Trainingsdatensatz stimmten 431 mit den Werten im Wissenschafts-Dokumentationssystem überein. Für die restlichen 37 wurden falsche Werte zugeordnet. Von 468 Attributwerten im Testdatensatz wurden 430 korrekt erkannt, bei 38 gab es fehlerhafte Extraktionen. Fehler sind vor allem bei den Attributen *Bluthochdruck*, *Familienanamnese*, *Fibroscan* und *Ultraschall* aufgetreten. Auffällig ist besonders die Fehlerhäufigkeit beim Attribut *Bluthochdruck*. Dieses Attribut wurde positiv beurteilt, wenn im Freitext der systolische Blutdruck den Wert 130 überschritten hat. Wie sich aber im Nachhinein herausstellte, war bei der manuellen Erfassung dieses Wertes folgende Regel angewandt worden: *Bluthochdruck* wurde diagnostiziert, falls der systolische Blutdruck 130 überstieg oder eine Medikation gegen Bluthochdruck vorlag oder eine arteriellen Hypertonie dokumentiert wurde. Die höheren Fehlerraten bei den anderen Attributen resultierten aus sprachlichen Variationen, die für die Wertegenerierung erkannt und zusammengeführt werden mussten, aufgrund der Verwendung nichtstandardisierter Text-Patterns. Ein Beispiel hierfür ist die Wertzuordnung für das Attribut *Ultraschall*, bei der verschiedene Varianten für das Vorliegen einer *Steatosis Hepatis* (Fettleber) interpretiert werden müssen z.B: „vereinbar mit“, „erhöht Echodichte i.S. einer“, „Hinweise auf“, „geringgr.“, „geringr.“. Die korrekte Erkennung dieser Werte erfordert anspruchsvollere NLP-Methoden als die reine Verwendung von regulären Ausdrücken. Im Vergleich zu diesen Werten ist die Extraktion von Attribut-Werte Paaren, die durchgängig quasi-standardisiert mit geringen Abweichungen im teilstrukturierten Dokumententemplate vorkommen, mit einer geringen Anzahl von regulären

[1]http://uima.apache.org/downloads/releaseDocs/2.3.0-incubating/docs/pdf/tutorials_and_users_guides.pdf

Ausdrücken und mit einer geringen Fehleranfälligkeit möglich. Eine Methodenübersicht für klinische Informationsextraktionssysteme basierend auf medizinischen Freitexten ist in [7] beschrieben wobei grundsätzlich zwischen regelbasierten Methoden und Methoden aus dem Bereich des maschinellen Lernens unterschieden werden kann.

	Trainingsdaten	Testdaten
Körpergröße	0.95	1.00
Gewicht	0.90	0.95
BMI	0.90	0.97
Bauchumfang	0.92	0.92
Hüftumfang	0.92	0.92
syst. Blutdruck	0.95	0.95
Typ II Diabetes	0.97	0.97
Familienanamnese	0.92	0.82
Bluthochdruck	0.69	0.74
Fibroscan	0.97	0.90
IQR	1.00	1.00
Ultraschall	0.95	0.85
	0.92	0.92

Tabelle 1: Ergebnisse der Informationsextraktion

5. Zusammenfassung und Diskussion

Ziel dieser Arbeit war es, die Genauigkeit der Transformation definierter Attribute aus EPAs in eine strukturierte Form zu beurteilen, unter Berücksichtigung der vorherrschenden Routinedokumentationsqualität in einem KIS. In unserem Beispiel wurde aufgezeigt, dass ca. 50% relevanter Studienattribute in einem teilstrukturierten Datenformat liegen. UIMA wurde unter Verwendung von regulären Ausdrücken gewählt, um relevante Attribute aus der EPA, speziell auf diese Texte optimiert, zu extrahieren. Die Erkennungsrate lag für den Trainings- und Testdatensatz bei über 0,9 (861/936). Können Attributwerte und deren Ausprägungen, die in einem quasi standardisierten Textformat vorkommen, mit einem minimalen Satz von regulären Ausdrücken und zufriedenstellend extrahiert werden, so ergeben sich zusammenfassend folgende Schwierigkeiten für die Informationsextraktion auf rein syntaktischer Ebene: Tippfehler: „Hüftumfang: 13 cm“, „Staeatosis hepatis“; Inkonsistenz: „Fibro-Scan“, „FIBROSCAN“, „Fibro Scan“, „Grösse von 1,55m“ versus „155 cm“; Redundanz: Mehrfachdokumentation desselben Attributes in verschiedenen Dokumenten (mit verschiedenen Ausprägungen); Sprachliche Ausdrucksvariationen: Pre-positive patterns: „geringgr.“, „ggr.“, „vereinbar mit“; Post-positive patterns: „ist vorhanden“; Pre-negative patterns: „kein Hinweis auf“; Post-negative patterns: „ist nicht vorhanden“.

Die Arbeit beinhaltet folgende Limitationen:

NLP: Die in dieser Arbeit beschriebene *Aggregierte Analysis Engine* (AAE) ist von einer vollständigen syntaktischen und semantischen Analyse des klinischen Freitextes mit NLP-Methoden weit entfernt. Es wurden reguläre Ausdrücke verwendet. Auch eine Gegenüberstellung zu maschinellen Lernmethoden wäre von Interesse. Eine manuelle Annotation eines Trainings- und Testkorpus ist dafür notwendig, die Erstellung zeitaufwendig, da die Trainingsdatenmenge hinreichend groß sein muss.

Kontext: Das System wurde ausschließlich auf syntaktischer Ebene evaluiert. Eine zu berücksichtigende Kontextgranularität reicht dabei von Satzebene (typisches Beispiel ist die Angabe eines Zielgewichts im medizinischen Freitext, das nicht dem aktuellen Gewicht entspricht), bis zur Dokumentenebene, da hier auch mehrere Dokumente desselben Typs vorliegen können und das richtige gewählt werden muss.

Speziell die Kontextdetektion stellt einen nächsten Schritt für die Verbesserung des Systems dar und könnte in Ahnlehnung an das System ConText [8] implementiert werden. ConText ist dabei ein System basierend auf regulären Ausdrücken, dass einem detektierten klinischen Zustand drei verschiedene kontextuelle Eigenschaften zuordnet und bewertet: Negation (Negated/Affirmed), Temporality (Historical/Recent/Hypothetical), und Patient Experience (Yes/No). Somit könnten Mehrfachdetektion wie z.B. aktuelles Gewicht versus Zielgewicht vermieden und Mehrdeutigkeiten aufgelöst werden.

Diese Arbeit wurde im Rahmen von GEN-AU III (GATiB II, Subprojekt Data Management and Biocomputing), gefördert vom Bundesministerium für Wissenschaft und Forschung, durchgeführt.

6. Referenzen

[1] J. Pathak, R.C. Kiefer, S.J. Bielinski, and C.G. Chute. Mining the Human Phenome using Semantic Web Technologies: A Case Study for Type 2 Diabetes. In AMIA Annual Symposium Proceedings, volume 2012, page 699-708, American Medical Informatics Association, 2012.

[2] T. Botsis, G. Hartvigsen, F. Chen, and C.Weng. Secondary use of EHR: data quality issues and informatics opportunities. AMIA Summits on Translational Science Proceedings, 2010:1, 2010.

[3] D. Segagni, V. Tibollo, A. Dagliati, L. Perinati, A. Zambelli, S. Priori, and R. Bellazzi. The onco-i2b2 project: integrating biobank information and clinical data to support translational research in oncology. Studies in health technology and informatics, 169: 887-891, 2011.

[4] H. Xu, Z. Fu, A. Shah, Y. Chen, N.B. Peterson, Q. Chen, S. Mani, M.A. Levy, Q. Dai, and J.C. Denny. Extracting and Integrating Data from Entire Electronic Health Records for Detecting Colorectal Cancer Cases. In AMIA Annual Symposium Proceedings, volume 2011, page 1564. American Medical Informatics Association, 2011.

[6] Ferrucci David and Lally Adam. Uima: an architectural approach to unstructured information processing in the corporate research environment. Natural Language Engineering, 10(3-4):327–348, 2004.

[7] Savova G., Kipper-Schuler K., Buntrock J. and Chute C. Uima-based clinical information extraction system. Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP, 39, 2008.

[8] Meystre S.M., Savova G.K., Kipper-Schuler K.C., Hurdle JF and others . Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform, 35:128–44, 2008.

[9] Harkema Henk, Dowling John N, Thornblade Tyler and Chapman Wendy W. Context: An algorithm for determining negation, experiencer, and temporal status from clinical reports. Journal of biomedical informatics, 42 (5):839–851, 2009.

Corresponding Author
Markus Kreuzthaler
Institut für Medizinische Informatik, Statistik und Dokumentation
Medizinische Universität Graz
Auenbruggerplatz 2, 8036 Graz, Österreich
Email: markus.kreuzthaler@medunigraz.at