# Extraktion kodierter Daten aus textuellen Befundberichten: Eine Fallstudie zu Echokardiographieberichten

Martin Toepfer[2], Philip-Daniel Beck[2], Georg Dietrich[2], Maximilian Ertl[1,2], Georg Fette[1,2], Peter Kluegl[1,2], Stefan Störk[1], Frank Puppe[2]

[1] Comprehensive Heart Failure Center, 97078 Würzburg
{georg.fette, pkluegl}@uni-wuerzburg.de, {ertl_m, stoerk_s}@ukw.de

[2] Universität Würzburg, Lehrstuhl Informatik VI, 97074 Würzburg
{philip.beck, dietrich, toepfer, puppe}@uni-wuerzburg.de

## 1 Einleitung

Im medizinischen Routinebetrieb fallen bei den verschiedensten Vorgängen große Menge an Daten an, die in den klinischen Informationssystemen (KIS) archiviert werden. Diese Datensenken sind ein reichhaltiger Fundus für die klinische Forschung um sowohl retrospektive Untersuchungen zu unterstützen, als auch für das Identifizieren geeigneter Kohorten für zukünftige Studien. Das Explorieren des Anteils der Freitexte dieser Daten gestaltet sich jedoch schwierig, da - abgesehen von einer unzureichenden Volltextsuche - Freitexte schlecht in die strukturierten Suchkonzepte von modernen Datenbanksystemen integriert werden können. Am Klinikum Würzburg wurde ein Informations-Extraktionssystem (IE) entwickelt, das aus den vorhandenen Freitexten wie z.B. Arztbriefen, Echokardiographieberichten, usw. strukturierte Daten gewinnen kann, die dann in gewohnter Manier von einer Datenbank gespeichert und dort abgefragt werden können. Die angewandte IE wird am Beispiel der Echokardiographieberichte illustriert.

## 2 Methoden

Das entwickelte IE-System wurde in Java implementiert und verwendet als grundlegende Datenverarbeitungsarchitektur das UIMA-Framework [1], das ideale Unterstützung für das Analysieren, Annotieren, Bearbeiten und Speichern von textuellen Daten bietet. Der Algorithmus des Systems erkennt die Konzepte durch Wortabgleich mit den Einträgen der Terminologie. Mit Hilfe syntaktisch erkannter Segmente und der Struktur der Terminologie werden daraufhin Mehrdeutigkeiten aufgelöst und Relationen zwischen diesen Konzepten identifiziert.

Die Vorgehensweise wird im Folgenden genauer erläutert (siehe auch [2]): Der Text wird zunächst in syntaktische Abschnitte unterteilt. Dabei wird ein einfacher regelbasierter Segmenttrenner verwendet, welcher an die Domäne angepasst wurde, da existierende Implementierungen keine zufriedenstellenden Ergebnisse in den medizinischen Teildomänen liefern. Die Zerlegung basiert größtenteils auf Satzzeichen (Punkte, Kommas, etc.), wobei Abkürzungen und Aufzählungen separat erkannt und bei der Segmentierung berücksichtigt werden. Da vorhandene medizinische Ontologien keine ausreichend feingranulare Abdeckung besitzen, wird eine auf die Textdokumente angepasste baumartige Terminologie verwendet, welche die möglichen medizinischen Konzepte der Domäne und deren Relationen spezifiziert. Für den initialen Textabgleich werden die Konzepte der Terminologie mit Schlüsselwörtern und Synonymen erweitert. Besitzen zwei oder mehr Konzepte dasselbe Schlüsselwort, so dass mehr als ein Konzept für eine betreffende Textstelle in Frage kommt, dann wird diese Mehrdeutigkeit durch eine Kontextsuche aufgelöst. Die Umgebung im Text, gegeben durch die Segmente und die darin erkannten Konzepte, in Kombination mit den entsprechenden

Relationen in der Terminologie liefert Aufschluss über die exakte Bedeutung einer mehrdeutigen Textstelle. Dieser Prozess kann mehrere Iterationen durchlaufen bis alle Textstellen, die mit Schlüsselwörtern übereinstimmen, einem Terminologiekonzept zugeordnet wurden. Die Erkennung negierter Konzepte ist hierbei bereits inhärent gelöst, da die Relationen in der Terminologie auch Negationen umfassen. Der gesamte Prozess läuft über eine grafische Benutzeroberfläche, mit welcher komfortabel während der Entwicklung der Terminologie ständig die Qualität der extrahierten Informationen überprüft werden kann.

## 3 Ergebnisse

Im Datenbestand des Klinikums Würzburg befinden sich ca. 21700 textuelle Echokardiographie-befunde aus den Jahren 2012-2013, die ca. 35 MB Speicherplatz umfassen. Ein durchschnittlicher Befundbericht enthält ca. 50 Informationen, von denen ein Teil numerisch und ein etwas größerer Teil qualitativer Natur ist. Dazu wurde manuell eine Terminologie bestehend aus 486 Knoten in einer Objekt-Attribut-Wert Struktur mit Synonymen und Kontextbezügen erstellt. Zur Evaluation wurde eine Teilmenge von 200 Dokumenten zufällig ausgewählt. Die Extraktion dauerte ca. 45 Sekunden. Der F-Score der Extraktionsqualität lag bei 96,2%, mit 93,2% Recall und 99,4% Precision. Der geringere Recall wird hauptsächlich von fehlenden Einträgen in der Terminologie verursacht. Diese lässt sich jedoch mit mehr Zeitaufwand einfach ergänzen, um die Extraktionsqualität weiter zu verbessern.

## 4 Diskussion

Mit dem dargestellten IE-System ist es möglich aus unstrukturierten Freitexten strukturierte, kodierte Information zu extrahieren. Die extrahierten Informationen werden im Rahmen eines Projektes des Deutschen Zentrums für Herzinsuffizienz am Klinikum Würzburg in ein Data Warehouse überführt, in welchem sie gemeinsam mit anderen strukturierten Daten für klinische Studien homogen abgefragt werden können. Im Gegensatz zu anderen IE-Systemen wie z.B. cTAKES [3], welche auf überwacht gelernten Modellen basieren, benötigt unser Ansatz keine annotierten Texte und kann somit bereits während der Terminologieerstellung eingesetzt werden. Weiterhin kann unser System nicht nur Begriffe aus einem Schlüsselwortkatalog identifizieren und kodieren, sondern auch feingranulare und problematische Mehrdeutigkeiten auflösen sowie den gefundenen Konzepten die in den Texten enthaltenen Eigenschaften (z.B. Adjektive, Zahlenwerte, etc.) zuordnen. Mit der dazugehörigen grafischen Entwicklungsumgebung für die IE-Terminologie ist es möglich, für eine gewünschte Domäne, wie in unserer Fallstudie Echokardiographien, in kurzer Zeit eine zufriedenstellende Extraktionskomponente zu entwickeln.

Diese Arbeit wurde Unterstützt durch die Förderung des Bundesministeriums für Bildung und Forschung (BMBF01 EO1004)

## 5 Referenzen

[1] Ferrucci D, Lally A: UIMA: An architectural approach to unstructured information processing in the corporate research environment. Natural Language Engineering 2004 10(3-4):327-48.

[2] Fette G, Ertl M, Wörner A, Kluegl P, Störk S, Puppe F: Information extraction from unstructured electronic health records and integration into a data warehouse. Lecture Notes in Informatics (LNI) 2012: 1237-51.
[3] Savova G, Masanz J, Ogren P, et al.: Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications; J Am Med Inform Assoc 2010 17(5):507-13.