

Die K-Box, eine komponentenbasierte Middleware-Lösung zum Austausch von medizinischen Dokumenten mit integrierter CDA-Erzeugung mittels Mustererkennung

Doan MH, Václavík M, Lott PL, Ückert F

Institut für Medizinische Informatik und Biomathematik, Universitätsklinikum Münster, Deutschland
doan@imfl.de

Einleitung und Fragestellung Die institutions- und sektorübergreifende Versorgung eines Patienten erfordert den Austausch von Dokumenten in Form von *Arztbriefen* [1]. Die Bedeutung dieser Dokumentart einerseits und das Fehlen eines einheitlichen, breit akzeptierten Formats andererseits treiben seit Jahren Standardisierungsbestrebungen auf diesem Gebiet an [2]. Ein international anerkannter ANSI-Standard für die Strukturierung und den Austausch von Arztbriefen ist die auf XML basierende *Clinical Document Architecture* (CDA) der Health Level Seven Inc. [3]. Die Patientendaten eines Arztbriefes liegen in elektronischer Form im Informationssystem des jeweiligen Versenders vor. Ein elektronischer Austausch von Arztbriefen im einheitlichen CDA-Format bietet Vorteile für Versorger und Patienten im Sinne der Integrierten Versorgung. Gegenwärtig ist ein elektronischer, institutions- und sektorübergreifender Arztbriefversand wegen der Heterogenität und den fehlenden Schnittstellen der eingesetzten Informationssysteme schwierig oder nicht möglich. Eine weitere Kommunikationsbarriere ist die Vielzahl an proprietären, unstrukturierten Arztbriefformaten der jeweiligen Institutionen, die eine automatische Weiterverarbeitung bei dem Empfänger verhindert. Die K-Box ist eine Middleware-Lösung für den elektronischen Austausch von medizinischen Dokumenten, insbesondere von Arztbriefen im CDA-Format. Sie besteht aus den zwei Teilsystemen *Strukturierung* und *Kommunikation*. Das Kommunikationsmodul stellt Dienste zur Verfügung um proprietäre Systeme mit der K-Box zu verbinden. Das Strukturierungsmodul konvertiert empfangene, unstrukturierte Arztbriefe mittels Mustererkennungsverfahren in das CDA-Format.

Material und Methoden Die *Informationsextraktion* ist eine Teildisziplin der Mustererkennung, die sich auf die Verarbeitung von Texten beschränkt. Das Ziel der Informationsextraktion ist die Erkennung von relevanten Informationen aus freien Texten bei gleichzeitigem „Überlesen“ irrelevanter Information. Ein verwandtes Gebiet der Informationsextraktion ist die *Textklassifikation*, dessen Aufgabe darin besteht, freie Texte zu analysieren und erkannte Muster in Form von Textpassagen mit einem entsprechenden Attribut zu markieren. Welche Methodik zur Textklassifikation ausgewählt werden soll, hängt von dem *Strukturierungsgrad* des Ausgangstextes ab. Allgemein wird zwischen *natural language processing* (NLP) und Nicht-NLP-Verfahren unterschieden. NLP-Verfahren versuchen durch die Analyse des Satzbaus unter Beachtung grammatikalischer Regeln und mit Hilfe einer Wissensbasis das menschliche Textverstehen zu imitieren. Sie sind dadurch zwar in der Lage aus unbekanntem Texten semantische Informationen zu extrahieren und diese zur Klassifikation heranzuziehen, ihre Implementierung ist aber extrem komplex. Nicht-NLP-Verfahren klassifizieren Texte mit Hilfe von statistischen Merkmalen (z.B. Analyse der Wortfrequenzen in Texten), sind weniger komplex und lassen sich schnell und einfach durch maschinelles Lernen an unterschiedliche Texte anpassen [4]. Die zugrunde liegende Architektur der K-Box ist offen und modular. Sie erlaubt die Implementierung, Verwendung und Kombination von verschiedenen Erkennungsverfahren, die je nach Format und Kontext des Eingangsdokuments unterschiedlich gute Ergebnisse liefern können.

Ergebnisse Die K-Box basiert auf *Java 2 Enterprise Edition* (J2EE)-Technologien und besteht im Kern aus einer J2EE-Anwendung (*Enterprise Application*). Die Komponenten sind *Enterprise Java Beans* (EJBs) vom Typ *Message Driven Bean* (MDB). Diese untergliedern sich in Kommunikations- und Strukturierungs- bzw. Transformationskomponenten. Dokumente werden von Kommunikationskomponenten empfangen und versendet. Zwischen Empfang und Versand wird das Dokument durch Transformationskomponenten in einer vorher definierten Reihenfolge sukzessiv verarbeitet. Neben Kommunikations- und Transformationskomponenten existieren Hilfskomponenten zur Verschlüsselung oder Pseudonymisierung von Inhalten. In welcher Reihenfolge ein Dokument durch welche Komponenten verarbeitet werden soll, wird von einem *Workflow* festgelegt. Dieser wird mit dem graphischen Werkzeug *Workflow-Designer* erstellt. Die Konfiguration und Verwaltung der K-Box geschehen über eine Web-Applikation. Die Komponentenverwaltung bietet die Möglichkeit, neue Komponenten von der *K-Box-Community* herunterzuladen und zu installieren. Die K-Box-Community ist ein Entwicklerportal zur verteilten Entwicklung von K-Box-Komponenten. Zum Hauptbestandteil der K-Box gehören Informationsextraktionskomponenten. Die adaptive Strukturierungskomponente verarbeitet nach einer Trainingsphase beliebige Word- und RTF-Dokumente und wandelt sie in CDA Level 2 konforme Dokumente um. Dabei werden die einzelnen Textblöcke inhaltlich klassifiziert und mit entsprechenden LOINC-Codes (für die spätere maschinelle Weiterverarbeitung des Dokuments) versehen. Die Dokumententransformation erfolgt dabei in drei Schritten: Klartextextraktion aus dem proprietären Dokumentenformat, adaptive Informationsextraktion (CDA-Headerdaten wie Patient und Autor) und Textblockklassifikation.

Diskussion Mit der K-Box sollen möglichst viele Versorger miteinander integriert werden. Ein besonderes Ziel des Projekts ist es, auch kleineren Versorgern mit eingeschränktem IT-Budget eine Möglichkeit zum elektronischen Arztbriefaustausch anzubieten. Die K-Box lässt sich mit einer angepassten Konfiguration und einer entsprechenden Bestückung von geeigneten Komponenten flexibel auf einen konkreten Bedarf des Anwenders anpassen. So kann sie sowohl in Krankenhäusern als auch in Arztpraxen eingesetzt werden. Die K-Box fungiert als Kommunikationsdrehscheibe und verbindet Praxisverwaltungssysteme mit Krankenhausinformationssystemen, die bisher zueinander inkompatibel waren. Die Freigabe der Komponentenentwicklung an die öffentliche Community wird die Anzahl der verfügbaren Komponenten und die Unterstützung von proprietären Systemen erhöhen. Eine Testversion der K-Box wurde in der Thorax-, Herz- und Gefäßchirurgie des Universitätsklinikums Münster installiert. Dort dient die K-Box zur Konvertierung der Entlassungsbriefe in PDF-Dokumente, die anschließend mittels PKI-Verfahren (*Public Key Infrastructure*) verschlüsselt und als Email-Anhang an die adressierten Empfänger verschickt werden. Die Evaluation der Testergebnisse geschieht in einer noch nicht abgeschlossenen Dissertationsarbeit am Institut für Medizinische Informatik und Biomathematik des gleichen Hauses. Am selben Institut werden ebenfalls im Rahmen einer noch nicht abgeschlossenen Diplomarbeit verschiedene Informationsextraktionsverfahren entwickelt und evaluiert. Die Entwicklung der K-Box erreichte zum Zeitpunkt dieser Veröffentlichung den zweiten von vier geplanten Meilensteine. Der Abschluss der ersten fertigen Version ist für das Ende des Jahres 2006 geplant.

Literatur

- [1] F. Warda, G. Noelle: *Telemmedizin und eHealth in Deutschland: Materialien und Empfehlungen für eine nationale Telematikplattform*. DIMDI, Deutsches Institut für medizinische Dokumentation und Information, 1. Auflage, 2002, ISBN 3-89906-701-0: 129.
- [2] Australian Standard™. Implementation of Health Level Seven (HL7) Version 2.3.1. Part 6: Referral and discharge summary, <http://www.hl7.org.au/Docs/AS4700.6-4004%20V2.3.1%20-%20Referral%20and%20discharge%20summary.pdf>.
- [3] R. H. Dolin, L. Alschuler, C. Beebe, P. V. Biron: The HL7 clinical document architecture. In: *Journal of the American Medical Informatics Association*, 8-6, 2001: 552ff.
- [4] R. O. Duda, P. E. Hart, D. G. Stork: *Pattern Classification*, 2nd Edition, New York: John Wiley & Sons Inc.; 2001.