Informationsextraktion aus medizinischen Texten basierend auf einer multiaxialen Indexierung

Denecke K, Kohlhof I
ID Information und Dokumentation im Gesundheitswesen GmbH, Platz vor dem neuen Tor 2, 10115 Berlin
K.Denecke@id-berlin.de

1. Fragestellung
Medizinische Dokumentation ist ein wichtiger Begleiter klinischer Behandlungsprozesse. In Befunden, Entlassungsberichten oder Arztbriefen werden Beobachtungen, Folgerungen, Diagnosen und Prozeduren überwiegend in Form von Freitext dokumentiert. Um zu verhindern, dass diese Informationen immer wieder neu aufgenommen und dokumentiert werden müssen, ist eine Wiederverwendung von bereits Erfasstem sinnvoll und erstrebenswert. Aber wie und inwieweit können ganz gezielt automatisch Informationen aus einem Text herausgelesen und weiterverarbeitet werden?
In diesem Beitrag wird ein Verfahren zur automatischen Extraktion von Informationen aus medizinischen Texten vorgestellt. Grundlage dafür bildet eine semantische Repräsentation der relevanten Inhalte eines Textes, die mittels einer mulitaxialen medizinischen Nomenklatur automatisch erstellt wird.

2. Ausgangsmaterial
Startmaterial für die Entwicklung eines Verfahrens zur Informationsextraktion sind von unterschiedlichen Ärzten einer Station verfasste chirurgische Arztbriefe eines Krankenhauses. Die Briefe enthalten kaum Schreibfehler, es handelt sich um grammatikalisch vollständige Sätze oder Nominalphrasen. Die einzelnen Paragraphen werden in der Regel einheitlich durch wiederkehrende Phrasen eingeleitet (z.B. „Die stationäre Aufnahme erfolgte.....“, „Wir entließen....“, „Der postoperative Verlauf gestaltete sich....“). Diese Phrasen bzw. daraus abgeleitete Triggerworte werden zur automatischen Klassifizierung der inhaltlichen Abschnitte eines Briefes herangezogen.

3. Erzeugen einer semantischen Repräsentation
Bevor Informationen aus einem medizinischen Text extrahiert werden können, muss dieser in eine standardisierte, maschinenlesbare Form überführt werden. Dazu wird das in [5] vorgestellte Verfahren verwendet, das mittels einer multiaxialen Nomenklatur (Wingert Nomenklatur [2]) und bereits bestehenden Sprachtechnologien (morphologische Analyse basierend auf einem konzeptbasierten Morphemlexikon, Indexierungsalgorithmus für Nominalphrasen [3,4], eingesetzt in den Produkten der Firma ID Berlin (http://www.id-berlin.de)) automatisch eine semantische Repräsentation medizinisch relevanter Inhalte eines Textes in Form von konzeptuellen Graphen [1] erzeugt. Die verwendete Nomenklatur besteht aus den 10 Achsen Diagnose (D), Morphologie (M), Funktion (F), Prozedur (P), Verfahren (V), Topografie (T), Job (J), Stoffe (W), Ätiologie (E) und Info (G). Die Verarbeitung erfolgt in fünf Schritten: Zunächst wird ein zu analysierender Text anhand von Absatzmarken in Absätze zerlegt. Die erkannten Absätze werden anhand von Triggerworten (z.B. „stationäre Aufnahme“) einer der zur Verfügung stehenden Klassen (z.B. Paraklinik, Anamnese, Aufnahmegrund, Behandlung etc.) zugeordnet. Die Paragraphen werden dann weiter in Sätze zerlegt, anschließend wird jeder Satz in Segmente zerlegt (z.B. [Thorax-CT] [mit Kontrastmittel] [im Juni 2005]). Ein Segment ist zum einen die Teilphrase vom Beginn des Satzes bis zur ersten Präposition, zum anderen die Teilphrasen, die jeweils mit einer Präposition beginnen und bis zur nächsten Präposition oder zum Satzende reichen. In einem nächsten Schritt werden Zahlenausdrücke (z.B. Oktober 2005, Pantoprazol 1-0-1) sowie in medizinischen Texten häufig wiederkehrende Phrasen (formelhafte Ausdrücke, wie z.B. Verdacht auf, Zustand nach) extrahiert und gesondert verarbeitet. Die einzelnen Segmente werden anschließend auf Indexketten über der Nomenklatur abgebildet. Pro Index wird ein semantisches Konzept erzeugt, das neben dem Index, die dazugehörige Beschreibung sowie ggf. die semantische Rolle umfasst. Mit Hilfe der Achseninformation, die in jedem Index enthalten ist (z.B. Index T000B32 (Lunge) liefert die Information, dass es sich um eine Angabe der Topographie handelt) wird die Hauptinformation pro Segment ermittelt. Der Ansatz geht davon aus, dass ein medizinischer Satz / Phrase entweder von einer Diagnose bzw. morphologischen Veränderung oder einer Prozedur bzw. Verfahren handelt. Diese Information wird als Hauptinformation eines Satzes betrachtet. Alle anderen Informationen werden als Attribute interpretiert, die diese Information spezifizieren. Die einzelnen Konzepte sowie die Abhängigkeiten zwischen diesen werden schließlich in einem Konzeptgraphen gespeichert. Abbildung 1 zeigt den Konzeptgraphen zu dem Satz „Thorax-CT mit Kontrastmittel im Juni 2005“. „CT“ ist die Hauptinformation des Satzes, die durch die anderen Informationen spezifiziert wird.

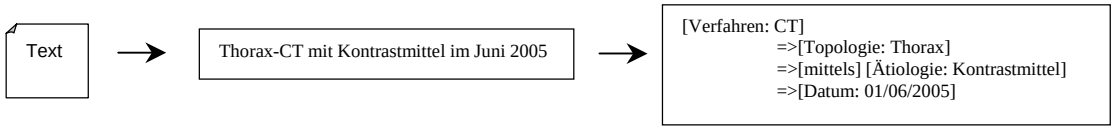

Abb. 1: Konzeptgraph zu der Phrase „Thorax-CT mit Kontrastmittel im Juni 2005“

4. Gezielte Suche nach bestimmten Informationen
Die semantische Darstellung zu einem Text soll nun dazu verwendet werden, ganz bestimmte Informationen aufzuspüren. Als zu extrahierende Informationen (genannt Templates) werden zunächst der Aufnahmegrund, Zustand bei Entlassung, Entlassungsmedikation, Diagnosen und Prozeduren betrachtet. Zu jedem Template sind verschiedene Slots definiert, in die extrahierte Informationen eingefügt werden sollen (z.B. sind die Slots Aufnahmeart, -anlass und –grund zu dem Template „Aufnahmegrund“ zu füllen).
Einige Informationen können direkt aus dem Originaltext extrahiert werden. Dazu wird mit Hilfe von regulären Ausdrücken nach bestimmten Wörtern / Zeichenketten gesucht. Das Vorhandensein oder Nicht-Vorhandensein eines Triggerwortes löst eine entsprechende Verarbeitung aus. Der Slot „Aufnahmeanlass“ kann z.B. den Wert „Notfallaufnahme“ haben. Taucht in dem untersuchten Abschnitt das Triggerwort „Notfall“ (bzw. ein Wort, das dieses Triggerwort enthält) auf, wird der Slot mit dem Wert „Notfallaufnahme“ gefüllt. Auch die Bezeichnung eines Abschnittes, in dem ein Triggerwort erkannt wird, kann relevant sein, da daraus zusätzliche Informationen abgeleitet werden können (z.B. tritt eine Diagnose im Abschnitt Aufnahme auf, wird sie als Aufnahmediagnose gekennzeichnet).
Andere Informationen können aus Konzeptgraphen extrahiert werden. Dazu wird zunächst zu relevanten Abschnitten mit dem oben beschriebenen Verfahren eine semantische Repräsentation erzeugt. Aus den Konzeptgraphen werden dann die gesuchten Informationen ermittelt. Dies umfasst die Ermittlung von ganz bestimmten Konzepten (z.B. Konzept für „beidseits“) oder Konzepten eines bestimmten Typs (z.B. vom Typ Diagnose) oder mit einer bestimmten semantischen Rolle (z.B. Zustand_nach) aus einem relevanten Konzeptgraphen. Viele gesuchte Informationen liegen in dem erzeugten Konzeptgraphen bereits vor und müssen nur den entsprechenden Slots zugewiesen werden. Soll z.B. aus dem Abschnitt Aufnahme der Aufnahmegrund extrahiert werden, wird der relevante Abschnitt zunächst semantisch repräsentiert. Der erzeugte Konzeptgraph wird nach Diagnosen bzw. morphologischen Veränderungen (Index aus der Achse Diagnose, Morphologie oder Funktion) durchsucht. Teil-Konzeptgraphen, die das gesuchte Konzept enthalten, werden in die Zielstruktur überführt. Abbildung 1 zeigt für einen Abschnitt Aufnahme die semantische Repräsentation sowie das damit gefüllte Template Aufnahmegrund.

Abschnitt Aufnahme:	|-- Satz (Aufgetreten waren in den Stunden zuvor....)
	  |-- (Verfahren) Erbrechen
	  |-- (Info) Stunde
	  |-- (Info) mehrfach
	  |-- (Info) vorher
	  |-- (Funktion) Bauchschmerzen
	  |-- (Info) rechtsbetont
		Extrahierter Aufnahmegrund
		
		Datum der Feststellung : 03/11/2005
		Diagnosetyp : Aufnahmediagnose
		Diagnosebezeichnung: Bauchschmerzen
		Verlässlichkeit: gesichert
		Seitenlokalisation: rechtsbetont

Der Patient stellte sich im Beisein seiner Mutter am Nachmittag des 03.11.05 in der Rettungs-stelle unseres Hauses vor. Aufgetreten waren in den Stunden zuvor mehrfaches Erbrechen sowie rechtsbetonte abdominelle Schmerzen.		

Abb. 1 Extraktion des Aufnahmegrundes aus dem Abschnitt Aufnahme: Der Abschnitt (linke Spalte) wird in einen Konzeptgraphen abgebildet (mittlere Spalte). Um das Template zum Aufnahmegrund zu füllen, werden sowohl der Eingabetext (Datum, Diagnosetyp, Verlässlichkeit) als auch der Konzeptgraph (Diagnosebezeichnung, Seitenlokalisation) verarbeitet und die gesuchten Informationen extrahiert.

5. Diskussion und Ausblick

Besonderheit und Vorteil des vorgestellten Systems ist die Verwendung von bereits existierenden und in der Praxis erfolgreich eingesetzten Werkzeugen zur morphologischen Analyse und Indexierung. Dadurch stehen wichtige Komponenten eines Systems zur Informationsextraktion bereits zur Verfügung. Aktuell wird das vorgestellte Verfahren evaluiert und erweitert.

Da das System zunächst für chirurgische Arztbriefe eines ausgewählten Krankenhauses erstellt wurde, wird ein Schwerpunkt weiterer Untersuchungen die Übertragung des Systems auf andere Texte (aus anderen Abteilungen, mit anderem Aufbau) sein. Triggerwörter und der Aufbau der Dokumente wurden manuell ermittelt und für die Erstellung der Extraktionsregeln genutzt. In Hinblick auf Anpassungen für andere Texte wird untersucht werden, ob Extraktionsregeln zum einen auch für unbekannte Dokumente verwendet werden können. Zum anderen wird geprüft, ob sie weitgehend automatisiert ermittelt werden können.

Das vorgestellte Verfahren ermöglicht es, gezielt Informationen aus einem natürlichsprachlichen Text zu extrahieren und in eine strukturierte Form zu überführen. Damit können schnell die relevanten Informationen aus einem Text ermittelt und weiteren Applikationen zur Verfügung gestellt werden.

6. Literatur

[1] Sowa J F: Knowledge Representation: Logical, Philosophical, and Computational Foundations. Brooks Cole Publishing Co., Pacific Grove, CA; 2000: 476-488

[2] Wingert F: SNOMED Manual. Berlin: Springer; 1984.

[3] Schulz S, Hahn U: Morpheme-based, cross-lingual indexing for medical document retrieval. International Journal of Med Inf 2000; 59(3) : 87-99.

[4] Hahn U, Honeck M, Piotrowski M, Schulz S: Subword Segmentation - Leveling out Morphological Varieties for Medical Document Retrieval. In: Proc of the 2001 AMIA Annual Symposium. Washington, 2001; 229-234.

[5] Denecke K, Kohlhof I, Bernauer J: Use of multiaxial indexing for information extraction from medical texts. To appear in: Proceedings of the Workshop on Foundations of Clinical Terminologies and Classifications; Romania, 2006