Finden und Zusammenfassen für Ärzte aus der Knochenmarktransplantation

Brigitte Endres-Niggemeyer[1], Bernd Hertenstein[2] und Carsten Ziegert[2]

[1] Fachhochschule Hannover, Fachbereich IK
Ricklinger Stadtweg 120
30459 Hannover
Brigitte.Endres-Niggemeyer@ik.fh-hannover.de

[2] Medizinische Hochschule Hannover, Abt. Hämatologie und Onkologie
30623 Hannover
Hertenstein.Bernd@mh-hannover.de
Carsten.Ziegert@ik.fh-hannover.de

http://summit-bmt.fh-hannover.de/

Abstract: SummIt-BMT, ein wissensbasiertes System zum Zusammenfassen aus dem WWW für Ärzte in der Knochenmarktransplantation (KMT), hat als zentrale Komponente eine KMT-Ontologie mit prädikatenlogischen Kontexten. Fragen und Antworten werden mit den Benutzern über eine szenariobasierte Schnittstelle ausgetauscht. Vom ausgefüllten Szenario leitet das System einen Satz von Queries für Medline und Google ab. In den Antworten werden die Medline-Abstracts möglichst durch Volltexte von Verlagsservern ersetzt. Die Antwortmengen werden dann durch ein Textpassagenretrieval auf die aussichtsreichen Passagen reduziert. Diese Passagen werden von Agenten zusammengefasst, die analog zu Strategien kompetenter Menschen vorgehen. Unser Grundsystem ist testfähig. An den Zusammenfassungsagenten arbeiten wir.

1. Einführung und Systemüberblick

Benutzern kann man das Sichten großer Antwortmengen auf Suchfragen weitgehend ersparen, indem man die Antwortdokumente zusammenfasst. Dies ist besonders wünschenswert in Umgebungen, wo es auf den neusten Stand der Kenntnis ankommt. In der Knochenmarktransplantation ist das so, denn es sind Therapien zu planen und durchzuführen, die für Patienten überlebenswichtig sind. Das System SummIt-BMT (Summarize It in Bone Marrow Transplantation) soll Ärzten durch kognitiv fundiertes Zusammenfassen [En98] aus dem WWW eine schnelle Informationsaufnahme ermöglichen.

SummIt-BMT hat als zentrale Komponente eine KMT-Ontologie. Den Systemablauf veranschaulicht Abb. 1: Benutzer geben ihren Informationsbedarf in ein strukturiertes Szenario ein. Sie ziehen dazu Begriffe aus der Ontologie heran. Aus dem Szenario werden Fragen an Suchmaschinen abgeleitet. Die Summit-BMT-Metasuchmaschine stößt Google an und sucht in Medline, der zentralen Literaturdatenbank der Medizin. Das Suchergebnis wird aufbereitet. Dabei werden Links zu Volltexten verfolgt und die Volltexte besorgt. Die beschafften Dokumente werden mit einem Schlüsselwortretrieval auf Passagen untersucht, in denen sich Suchkonzepte aus der Frage / Ontologie häufen. Diese Passagen werden zum Zusammenfassen vorgeschlagen. In ihnen werden die Aussagen syntaktisch analysiert. Die Systemagenten untersuchen sie. Lassen Aussagen sich mit einer semantischen Relation an die Frage anbinden, tragen also zur deren Beantwortung bei, werden sie in die Zusammenfassung aufgenommen, es sei denn, andere Agenten machen Hinderungsgründe geltend, z.B. Redundanz. Das Ergebnis der Zusammenfassung wird in das Frage/Antwort-Szenario integriert. Präsentiert werden Exzerpte

aus den Quelldokumenten. Mit einem Link vermitteln sie einen sofortigen Rückgriff auf die Quelle. SummIt-BMT ist anschließend zum nächsten Durchgang von Informationssuche und Zusammenfassung bereit.

Weitgehend realisiert sind die Ontologie, die szenariobasierte Benutzeroberfläche und das Grundsystem zu Websuche und Textpassagenretrieval. Die Zusammenfassungsagenten sind in Arbeit.

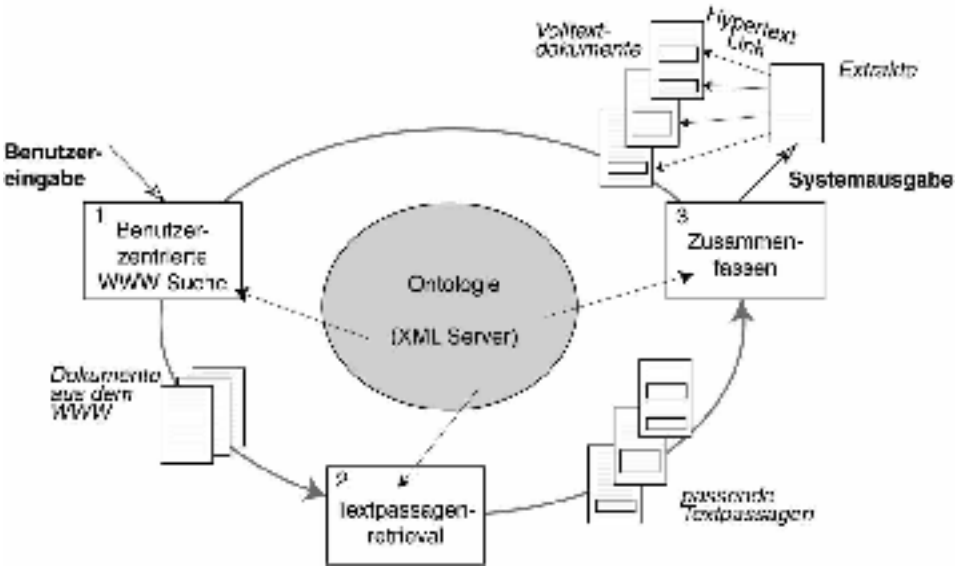

Abb. 1. Systemüberblick

2. Die Ontologie

Die korpusbasierte KMT-Ontologie umfasst gut 4400 Konzeptsätze. Sie wurden aus zwei Teilkorpora gewonnen: aus 9 Aufsätzen aus Blood (2000), einer zentralen Zeitschrift der KMT, und aus 6 Educational Papers (2000) der American Society for Hematology (ASH). Hinzu kamen und kommen noch in kleinen Mengen Begriffe aus den Benutzerszenarios und Ergänzungen von BearbeiterInnen. Konzeptsätze der Ontologie enthalten den Begriff selbst, Synonyme, eine Beschreibung und die Okkurrenzen des Konzeptes im Korpus. Sie wurden mit ihrem sprachlichen Kontext - meist einem Satz – aufgenommen, aus dem Kontext gelöst und aufgearbeitet. Nach der fachlichen Kontrolle entstehen daraus prädikatenlogische Kontextausdrücke. Aussagen bzw. aus ihnen aufgebaute Kontextausdrücke verbinden mehrere Konzepte der Ontologie. Abb. 2 zeigt einen Kontextausdruck und illustriert, wie er gebaut ist. Die wesentliche Idee von [MB93] ist es, Aussagen, die nur unter bestimmten Kontextbedingungen wahr sind, mit Aussagen über die Einschränkungen zu versehen. So entsteht ein Kontextausdruck, der in einer Prädikatenlogik erster Stufe wahr ist.

Die KMT-Ontologie hat zur Zeit gut 1400 Kontextausdrücke, für die rund 2800 Propositionen gebraucht werden. Benutzt werden sie zur Definition der Zusammenfassungsziele in Szenarios (s. Abschnitt 3) und von den Zusammenfassungsagenten (s. Abschnitt 5). Die Ontologie ist als Filemaker-Anwendung und als XML-Datenbank realisiert. Die Anpassung an den Topic-Maps – Standard XTM 1.0 ist geplant.

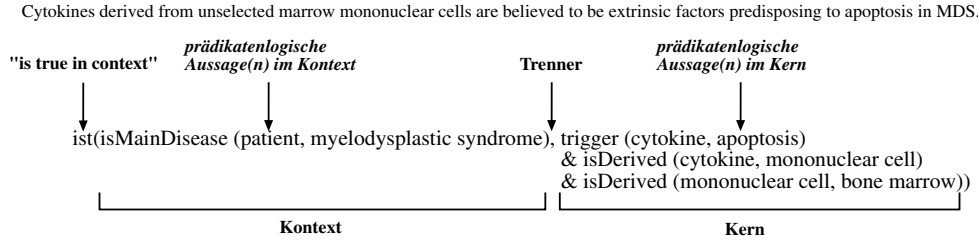

Abb. 2. Ein Kontextausdruck mit Kontextteil und Kern

3. Die Benutzeroberfläche

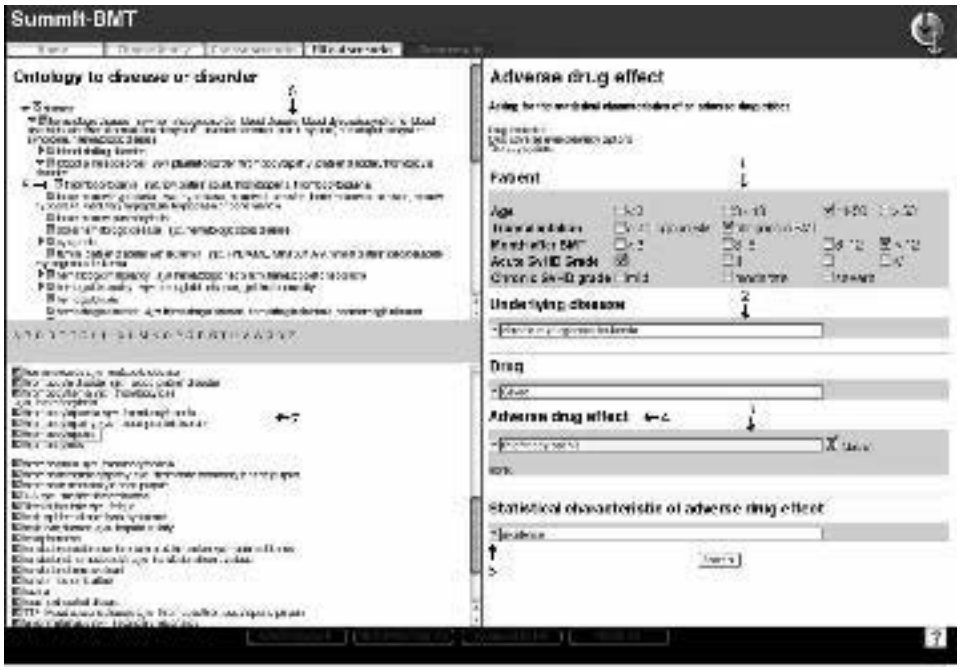

Abb. 3. Ein Szenario auf der rechten Seite des Bildschirms. Links die Ontologie (systematisch und alphabetisch)

Die Benutzeroberfläche von SummIt-BMT bietet Szenarioformulare an. Sie wurden empirisch aus Alltagssituationen von KMT-Ärzten entwickelt. Die bekannten Merkmale der Situation setzen den Rahmen des Szenarios, die Informationslücke definiert den Kern der Frage und das Thema der Zusammenfassung. Damit realisieren die Szenarios lange bekannte Ziele des benutzerorientierten Information Seeking ([MK98] u.a.). Die Szenarios strukturieren die Situationen des Informationsbedarfs. Sie erleichtern so die Frageformulierung und bauen unbefriedigenden Suchergebnissen vor. Sie sind in Familien geordnet und durch Links verbunden. Zur Zeit umfasst die Szenariobibliothek 41 voll ausgearbeitete Szenarios, die auf 131 Fragebeschreibungen von Ärzten beruhen. Die Szenarios sollen die gängigen Fragen bedienen. Ein Auffangszenario leitet Restfragen zu geeigneten Webadressen weiter – bisher zu PubMed und Google.

Abb. 3 zeigt die Benutzersicht auf ein Szenario. Rechts steht das eigentliche Szenario. Es behandelt Nebenwirkungen von Medikamenten. Unter dem Titel finden sich Links zu

verwandten Szenarios. Der Patient wird in einem Block von Merkmalen zum Anhaken beschrieben (1). Es gibt einfache Slots für Benutzereinträge (2) und solche, die Listen aufnehmen können (3). Jeder Slot ist mit einer Ontologieklasse vorbelegt. Sie wird durch Anklicken der Überschrift aufgerufen (4). Links auf dem Monitor erscheint die Ontologie, oben in systematischer Ordnung (6), unten alphabetisch (7). Ein kleines „D"-Icon (8) besorgt die Beschreibung eines Konzeptes. Aus der Ontologie lässt sich ein Konzept per Klick in den passenden Slot übertragen. Benutzer kennzeichnen ihre Frage, indem sie das Fragezeichen vor einem Feld (5) anschalten. Im Beispielfall verursacht Glivec einen Abfall der Thrombozyten. Der Arzt will die Inzidenz wissen.

Hinter dem Szenario vermittelt eine prädikatenlogische Interpretation (Abb. 4) zwischen der Benutzersicht und der Systemsicht. Sie verwendet prädikatenlogische Aussagen aus der Ontologie. Die Benutzereingaben werden durch Variable (mit # gekennzeichnet) eingebracht. Aus der Szenariointerpretation wird die Zielvorgabe für das Zusammenfassen abgeleitet (links in Abb. 4). Unberührte Slots werden ignoriert. Die Propositionen, die Angaben aus gefüllten Slots aufnehmen, wandern in den Kontext des Frage-Kontextausdruckes. Nur Aussagen, die Fragekonzepte enthalten, kommen in den Kern. Auch die Suchfragen werden abgeleitet. Sie enthalten Konstanten des Szenarios und Werte aus der Benutzereingabe.

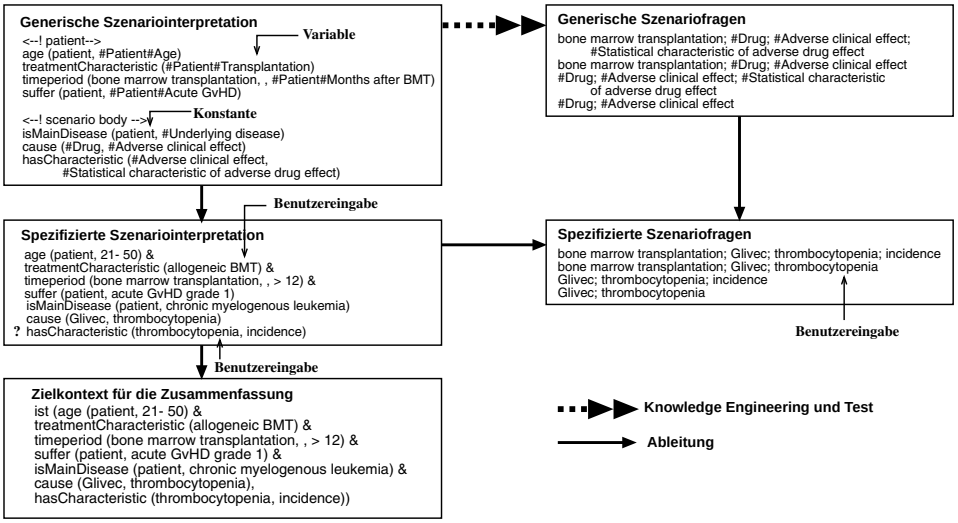

Abb. 4. Interpretation des Szenarios Adverse drug effect

4. Grundsystem zum Information Retrieval und Textpassagen-Retrieval

Die technische Grundlage des Gesamtsystems ist eine XML-Datenbank. Diese enthält die Ontologie mit Fachkonzepten und prädikatenlogischen Aussagen, eine prädikatenlogische Interpretation der Szenarios sowie Regeln zur Generierung der Suchanfragen aus den Szenario-Inhalten, Adapter für die Suche im WWW und Informationen zum Download von Zeitschriftenartikeln von Verlagsservern.

Information Retrieval. Beim Retrieval wird zunächst wird das ausgefüllte Szenario (s. Abschnitt 3) in XML-Darstellung vom Client (Web-Browser) an den Server übertragen. Mit der Identifikation des Szenarios werden aus der Datenbank die Regeln zur Generie-

rung der Suchanfragen ermittelt. Diese Regeln enthalten die konstanten Suchbegriffe aus dem Szenario und die retrievalrelevanten Angaben des Benutzers. Für jedes Szenario werden eine Mindestanzahl zu erwartender Resultate angegeben sowie Alternativregeln, die eine weniger restriktive Suchanfrage generieren (Relaxationsstrategie). Mit Hilfe der Ontologie wird eine resultierende Suchanfrage um Synonyme und alternative Schreibweisen erweitert. Solange die vorgegebene Mindestanzahl an Resultaten nicht erreicht ist und Alternativregeln zur Verfügung stehen, werden Suchanfragen generiert und an die Adapter für die WWW-Suche übergeben.

Adapter bestehen im Wesentlichen aus einer Basis-URL für die anzusprechende Suchmaschine (bisher Medline und Google) sowie aus Parsing-Regeln zur Verarbeitung des Suchergebnisses, das als HTML- bzw. XML-Seite angenommen wird.

In der Regel liefern Suchmaschinen als Resultat eine Liste von URLs; Medline liefert hingegen Abstracts und bibliografische Daten zu einem Fachartikel. Mit Hilfe der Zeitschriften- und Verlagsinformationen in der XML-Datenbank wird aus den bibliografischen Informationen eine URL zum Download des Artikels vom Verlagsserver generiert. Als Ergebnis liegt nun eine Liste von URLs (allgemein im WWW sowie auf Verlagsservern) vor, mit der die gesuchten Dokumente geholt werden.

Textpassagen-Retrieval. Im Textpassagenretrieval werden die einschlägigen Textpassagen eines jeden Dokuments ermittelt. Ein Absatz gilt als einschlägig, wenn die Suchbegriffe aus dem Szenario oder ihre Spezialisierungen in ausreichender Anzahl darin vorkommen. Mit Hilfe der Ontologie wird deren Liste sukzessive um alle Unterbegriffe beliebiger Tiefe im Ontologiebaum erweitert (transitive Hülle). Anschließend werden alle Synonyme und Alternativschreibweisen der Listenelemente hinzugefügt. In jedem Absatz werden nun die Vorkommen der Begriffe gezählt. Die Bewertung eines Absatzes ergibt sich aus dem Anteil der Suchwörter an den Nicht-Stoppwörtern. Positiv bewertete Absätze werden den Zusammenfassungsagenten übergeben.

5. Erste Zusammenfassungsagenten

Sprachorientierte Zusammenfassungsagenten bilden einschlägige Aussagen im Eingabetext auf Propositionen und Kontexte der Ontologie ab. Am weitesten entwickelt ist der aufgabenorientierte Informationsextraktionsagent TextToProposition. Er verwendet empirisch ermittelte Vorkommensbeschreibungen für Propositionen der Ontologie. Relevanzagenten ermitteln relevante Aussagen im Sinne der Frage. Vorbilder finden sich im SimSum-System [En98]. Der erste Relevanzagent von SummIt-BMT wird redundante Aussagen eliminieren.

6. Einige bibliographische Angaben

[En98] Endres-Niggemeyer, B.: Summarizing information. Springer, Berlin, 1998.

[MK98] Marchionini, G.; Komlodi, A.: Design of interfaces for information seeking. In ARIST 33, Annual Review of Information Science and Technology. Information Today, Medford NJ, 1998, S. 89-130.

[MB93] McCarthy, J.; Buvac, S.: Notes on formalizing context. Proceedings of IJCAI'93, Chambery, France, 1993, S. 555–560.
http://www-formal.stanford.edu/jmc/context3/context3.html.