ANWENDUNG UND EVALUIERUNG SEMANTISCHER RETRIEVALTECHNOLOGIEN AUF MEDIZINISCHE BEFUNDTEXTE

Faulstich LC[1], Müller F[1], Sander A[1], Kreuzthaler M[2], Kaiser S[2], Errath M[2]

Kurzfassung

Die Recherche auf medizinischen Freitexten, wie Arztbriefen und Befunden, ist von großer Bedeutung für Forschung, Qualitätssicherung und Abrechnung. Dabei sind Schreibvarianten, Schreibfehler, Homonyme und Synonyme, sowie ontologische Beziehungen zu berücksichtigen. Ein von ID entwickeltes Verfahren zur semantischen Recherche wurde vom IMI auf verschiedenen Testcorpora mit Hilfe einer neu entwickelten Testumgebung evaluiert. Das Verfahren erwies sich als konkurrenzfähig zu von Experten optimierten Anfragen.

1. Einleitung und Fragestellung

Die Recherche auf medizinischen Freitexten, wie zum Beispiel Arztbriefen und Befunden, ist von großer Bedeutung in der klinischen Forschung, Qualitätssicherung und Abrechnung. Die Gründe dafür liegen hauptsächlich im stetigen Zuwachs an Informationen zu einem Patienten, von denen nach wie vor ein beträchtlicher Teil in Form von Freitext vorhanden ist. Aufgrund der Einführung von Krankenhausinformationssystemen und medizinischen Archivsystemen in Verbindung mit automatischer Schrifterkennung (OCR) sind diese Informationen zunehmend als elektronische Texte verfügbar. Die Beschränkung auf strukturierte und codierte Daten ist dagegen in vielen Fällen nicht ausreichend, da diese nur einen kleinen Teil der tatsächlich vorhandenen Informationen abbilden. Zudem gehen durch Codierung oft wesentliche medizinische Aussagen verloren [1].

Die Eigenheiten medizinischer Freitexte stellen jedoch besondere Herausforderungen an die Volltext-Suche: Krankenakten werden aufgrund der mehrheitlich internen Verwendung wenig auf sprachliche Schwächen hin überprüft, und orthographische Mängel sind in der klinischen Routine an der Tagesordnung. Solange der Text verständlich ist, gibt es ärztlich keinen zwingenden Grund, sich mit der Sprache auseinanderzusetzen.

Als Vokabular dient eine stark lokal variierende Mischung von lateinischen, deutschen und englischen Begriffen und Abkürzungen. So können Befunde zu schwer lesbaren „Telegrammen“ mutieren, wie z.B.: „*Ca. 2x1cm, große ovaläre Verschattung im UF li. Rez. re. lat. frei, li. teiladhärent*“.

[1] ID Information und Dokumentation im Gesundheitswesen GmbH & Co. KGaA (ID)
[2] Institut für Medizinische Informatik, Statistik und Dokumentation, Medizinische Universität Graz (IMI)

Obwohl für einschlägig geschulte Medizinerinnen und Mediziner durchaus verständlich, ist der Text doch für die maschinelle Handhabung mit traditionellen IR-Methoden sehr problematisch. Manche Abkürzungen haben zwar eine eindeutige Expansion, eine große Anzahl davon ist jedoch homonym und daher nur in ihrem Kontext sicher auflösbar. So sind *ADS* (*Aufmerksamkeits-defizitsyndrom* oder *akutes Durchfallsyndrom*) bzw. *HWI* (*Harnwegsinfekt* oder *Hinterwandinfarkt*) homonyme medizinische Abkürzungen, *ca.* (*circa*), *Ca.* (*Karzinom*) und *Ca* (*Kalzium*) sind homonyme Abkürzungen aus Schriftsprache, Medizin und Chemie. Die inkonsistente Verwendung von Satzzeichen und alternierenden Schreibweisen erschweren die Situation weiter. Durch einen einzigen „kleinen Tippfehler", kann aus dem Satz „Geschwulst, ca. 3 cm groß", folgender Satz werden: „Geschwulst, ca, 3 cm groß", wodurch die Unterscheidung zwischen einem Karzinombefund, einer ungefähren Größenangabe oder einer Erwähnung von Kalzium stark erschwert wird [1].

Aus diesen Gründen ist die im Information Retrieval übliche Verschlagwortung mit Lemmatisierer und Stemming unzureichend. ID hat daher ein ontologiebasiertes Retrieval-Verfahren entwickelt, das auf anspruchsvollen linguistischen Methoden zur Textaufbereitung aufsetzt. Diese Lösung für die semantische Recherche wird auf medizinischen Freitexten an der Medizinischen Universität Graz mit einer eigens dafür entwickelten Testumgebung evaluiert.

2. Material und Methoden

Auf der Basis ihrer in der medizinischen Dokumentation bewährten Textaufbereitungsmethoden hat ID Verfahren zur automatischen semantischen Repräsentation von medizinischen Freitexten entwickelt [2,3]: die zu verarbeitenden Texte durchlaufen eine Pipeline von Verarbeitungsstufen, die zunächst die äußere Gliederung eines Textes in Abschnitte, Sätze und Wörter erkennen. Danach werden Abkürzungen expandiert und Rechtschreibfehler korrigiert. Zusammengesetzte Wörter werden nach einem auf den Arbeiten von Wingert [4,5] und Göttsche aufbauenden Verfahren in ihre Bestandteile zerlegt und auf Konzepte der Wingert-Nomenklatur [6] abgebildet. Homonyme werden kontextabhängig disambiguiert, indem Kollokationsbeziehungen in einem Referenzkorpus und semantische Beziehungen im Semantischen Netz berücksichtigt werden. So kann z.B. die Abkürzung „A." vor „carotis" automatisch zu „arteria" disambiguiert werden und die Abkürzung „HWI" im Zusammenhang mit „Vorhofflattern" zu „Hinterwandinfarkt" (statt „Harnwegsinfekt"). Dieses Verfahren ist in der klinischen Praxis bewährt, wurde aber bisher nicht wissenschaftlich evaluiert. Durch die beschriebenen Textaufbereitungsmethoden kann den Erwartungen der Ärzteschaft an die Berücksichtigung von relativierenden und mehrdeutigen Ausdrucksformen und von uneinheitlichen Abkürzungen besser entsprochen werden. Gegenüber der in [2,3,11] verwendeten Erkennung von Satzbestandteilen mit regulären Ausdrücken und einem einfachen Parser für Präpositionalphrasen kommt inzwischen ein mehrphasiger Chunk-Parser zum Einsatz. Der damit aufgebaute Syntaxbaum (s. *Abbildung 1*) stellt neben der Phrasenstruktur Angaben zu Textbausteinen, quantitativen Angaben und Negationen bereit, die bei der Abbildung in die Wingert-Nomenklatur berücksichtigt werden. Durch diese sogenannte *kompositionale Indexierung* konnte die Qualität der Textaufbereitung deutlich gesteigert werden.

Zusätzlich berücksichtigt die Suche auch taxonomische und meronymische (Teil-Ganzes) Beziehungen: eine Suche nach *Neubildungen* findet auch *Karzinome*, eine Recherche nach Befunden am *Verdauungstrakt* auch Befunde an *Magen* oder *Darm.* Um dies zu erreichen, werden Suchbegriffe als Konzepte der Wingert-Nomenklatur repräsentiert und durch Beziehungen aus dem ID MACS® - medical semantic network - expandiert.

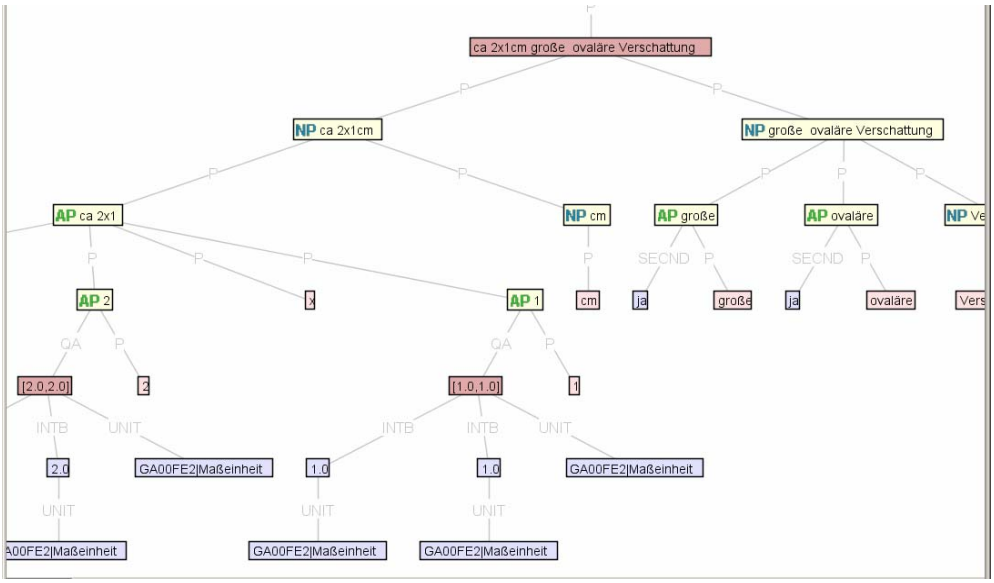

Abbildung 1: Vom Parser erzeugter Syntaxbaum eines Beispieltextes

3. Evaluierung

Zur Evaluierung dieses Verfahrens wurde vom IMI eine Testumgebung entwickelt, mit deren Hilfe es möglich ist, gleichzeitig sowohl eine durch eine lokale Expertin oder einen Experten formulierte expandierte Anfrage (siehe *Tabelle 1*, Spalte „Manuell expandierte Stichwortsuche“) als auch die semantische Recherche mit dem von ID zur Verfügung gestellten Information Retrieval Tool gegen einen *Goldstandard* – 3034 manuell beschlagwortete anonymisierte Befunde - durchzuführen und die entsprechenden Kennzahlen (Recall, Precision, Fallout) im Vergleich darzustellen. Somit ist ein Leistungsvergleich zwischen einer hochspezialisierten, manuellen Expertenabfrage und einem universellen semantischen Verfahren zur Textrecherche rasch und unmittelbar möglich.

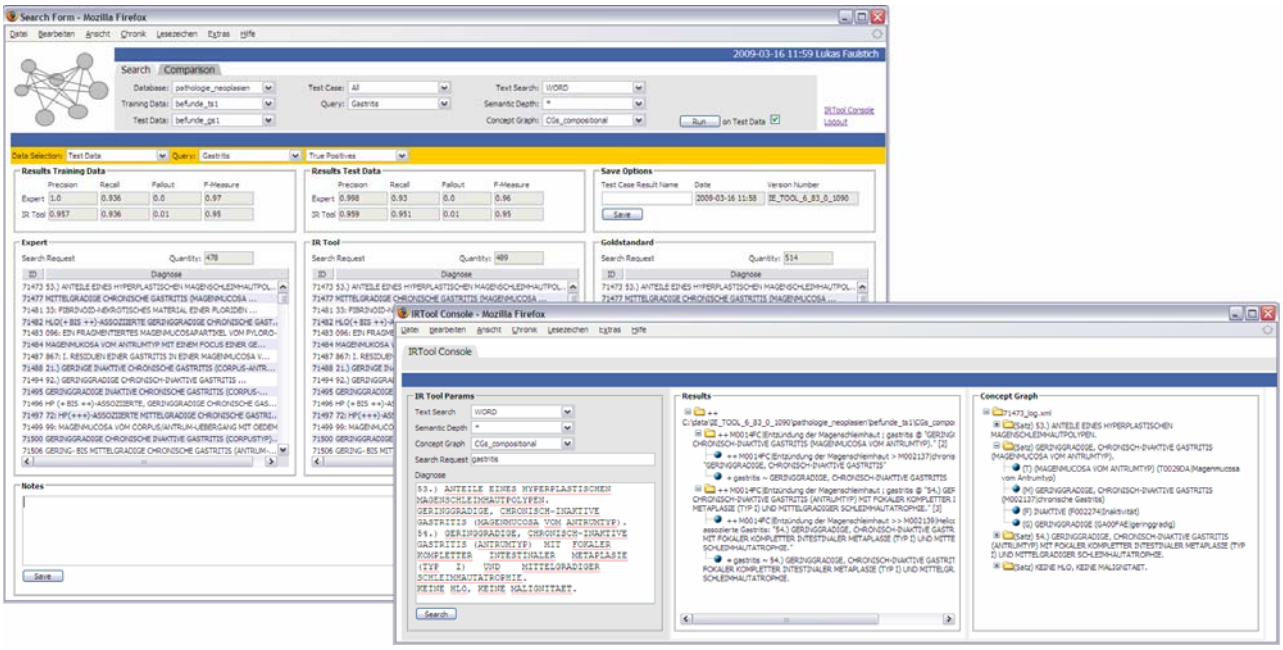

Abbildung 2: Screenshots der Testumgebung

Die Testumgebung ist als Web-Applikation ausgeführt (Apache Tomcat, MySQL, Spring Framework) wobei auf gute Übersichtlichkeit und einfache Benutzbarkeit besonderer Wert gelegt wurde.

Durch das Abspeichern aller Parameter und Ergebnisse in einer Datenbank können jederzeit Anfragen und Ergebnisse reproduziert und somit unterschiedliche Versionen des Information Retrieval Tools verglichen werden. Zusätzlich zu den Testdaten des Goldstandards existieren separate Trainingsdaten derselben Textsorte, die zur Entwicklung genutzt werden. Dadurch wird ein Lerneffekt ausgeschlossen, der sich lediglich auf die Spezifika des Goldstandards bezieht. Die Testdaten sind für die ID nicht einsehbar, sondern nur die resultierenden Kennzahlen daraus. Goldstandard, Trainings- und Testdaten sind je nach Verfügbarkeit beliebig und rasch austauschbar, um einen universellen Einsatz auf verschiedenen Befundkollektiven zu ermöglichen.

Konkret wurde die semantische Recherche auf einem Referenzkorpus von 3034 Texten mit mehreren Suchbegriffen getestet. Verglichen wurden die damit erzielten Ergebnisse mit einer naiven Stichwortsuche nach dem gegebenen Suchbegriff und einer manuell expandierten Stichwortsuche. Die semantische Recherche wurde anhand von Trainingsdaten (250 separate Befunde) für die Textsorte des Referenzcorpus optimiert. Die manuellen Anfrageexpansionen wurden dagegen von Experten iterativ auf dem Referenzkorpus optimiert (in der Praxis würden zusätzliche manuelle Nachbearbeitungen durch Experten folgen).

4. Ergebnisse

In *Tabelle 1* sind die Kennzahlen Precision (Prec) und Recall (Rec) für die naive Stichwortsuche, die manuell expandierte Stichwortsuche und für die Semantische Recherche (mit/ohne Einbeziehung von Eltern im semantischen Netz) für unterschiedliche Suchbegriffe gegen den Goldstandard dargestellt. Die Ergebnisse werden im Kapitel 5 diskutiert.

Tabelle 1: Datenbasis 3034 Befunde (Prec=Precision, Rec=Recall)

Suchbegriff	Anzahl	Naive Stichwortsuche		Manuell expandierte Stichwortsuche			Semantische Recherche			
							Ohne Eltern		Mit Eltern	
		Prec	Rec	Expansion	Prec	Rec	Prec	Rec	Prec	Rec
Gastritis	514	0,99	0,92	-	0,99	0,92	0,95	0,95	0,95	0,95
Hepatitis	20	0,86	0,65	%hepatitis%; %Nash%	0,86	0,65	0,88	0,80	0,17	0,85
Colitis	89	0,71	0,83	%colitis%; %kolitis%	0,71	0,83	0,70	0,82	0,70	0,82
Appendicitis	130	0,99	0,93	%appendi_itis%	0,99	0,95	0,98	0,97	0,62	0,95
Adeno-karzinom + Dickdarm	43	0,87	0,65	siehe Fußnote [1]	0,94	0,83	0,80	0,83	0,75	0,83
Adeno-karzinom + Colon	43	0,20	0,02	siehe Fußnote 3	0,94	0,83	0,50	0,09	0,72	0,67
Neubildung + Darm	213	0,92	0,17	siehe Fußnote [2]	0,75	0,91	0,73	0,92	0,73	0,76
Tumor + Prostata	34	0,46	0,35	%prostata%, (%ar_inom%"; %neoplasie%)	0,85	0,88	0,72	0,76	0,72	0,76

5. Diskussion

Eine Übersicht über Verfahren der Anfrageexpansion im Information Retrieval gibt [7]. Interaktive Anfrageexpansion mit Hilfe des UMLS-Metathesaurus wird in [8,9] beschrieben. Multilinguales Retrieval mit UMLS-Konzepten wurde im Projekt MUCHMORE [10] untersucht. Ein direkter Performance-Vergleich mit diesen Verfahren war nicht möglich, da die dort beschriebenen Forschungsprototypen nicht zur Verfügung standen.

Der größte Vorteil der semantischen Recherche liegt darin, dass für die Formulierung einer Anfrage keine Expertenkenntnisse und keine aufwändige Optimierung der Anfrage notwendig sind. Dies kommt besonders bei unspezifischen Suchbegriffen zum Tragen, da hier viele Unterbegriffe und Synonyme zu berücksichtigen sind und somit effektive manuelle Anfrageexpansionen schwierig sind (vgl. die Anfrage *Neubildung+Darm* in *Tabelle 1*). Weiter ist die semantische Recherche dort besonders nützlich, wo die Sprache innerhalb einer Textsammlung heterogen ist oder wo die Terminologien von Nutzern und Autoren voneinander abweichen. Die Auswertung zeigt, dass bei spezifischen Suchbegriffen (*Gastritis*) der Vorteil der semantischen Recherche gering ist. Bei komplexen Recherchen sind die Ergebnisse meist mit aufwändig durch Experten optimierten Anfragen vergleichbar und übertreffen in ihrer Ausbeute (Recall) die einfache Stichwortsuche deutlich. Ein weiterer Vorteil der semantischen Recherche ist das Erkennen und Ausfiltern von negierenden Aus-

[1] %adeno_ar_inom%,(%kolon%; %colon%; %rectum%; %rektum%; %dickd%; %sigm%; %asc%; %desc%; %trans%; %flex%)

[2] (%darm%; %colon%; %kolon%; %re_t%; %sigm%; %duoden%; %jejun%; %hemi_olektomie%; %ileum%; %ileo%; %appendix%; %coe_um%; %zoe_um%), (%Karzinom%; %Adenom%; %Kar_inoid%; %Lymphom%; %NHL%; %Sar_om%; %Myom%; %Neurom%); (%Tumor%, %neuroendokrin%);(%intraepithel%, %neoplasie%)

sagen (*Ausschluss von* etc.). Z.B. ist die einfache Stichwortsuche nach *Tumor+Prostata* sehr unbefriedigend. Manuell optimierte Anfrageexpansionen erfordern erfahrene und mit dem Korpus vertraute Experten (z.B. ist *%tumor%* wegzulassen, weil damit auch Vorkommen von *tumorfrei* gefunden würden). Manuell optimierte Anfragen haben gegenüber jeder automatischen Methode den grundsätzlichen Vorteil, dass sie viele Anpassungen bezüglich des zugrundeliegenden Korpus und der Intention des Benutzers erlauben. Obwohl noch sichtbare Defizite bestehen, erreicht die semantische Recherche im Vergleich zu manuell optimierten Anfragen bereits achtbare Ergebnisse und bietet einen deutlichen Mehrwert gegenüber einer naiven Stichwortsuche.

Folgende offene Probleme wurden identifiziert. Diskrepanzen zwischen prä- und postkoordinierter Repräsentation: eine Suche nach *Cholangitis* soll auch Befunde mit einer *Entzündung* in einem *Gallenweg* finden. Unterschiedliche Konzeptualisierungen beim Benutzer und in der Ontologie [11]: bei *Adenokarzinom+Colon* werden teilweise auch Vorkommen von anatomischen Oberstrukturen (*Dickdarm* statt *Colon*) erwartet. Wenn die Vorfahren erster Stufe im semantischen Netz hinzugenommen werden (vgl. Spalte „mit Eltern“ in *Tabelle 1*), so verbessert sich zwar der Recall in diesem und ähnlich gelagerten Fällen, allerdings zu dem Preis einer verringerten Präzision bei anderen Suchbegriffen (siehe z.B. *Hepatitis*). Daher müssen erst noch Kriterien für das gezielte Einbeziehen von Vorfahren gefunden werden.

6. Literaturangaben

[1] HOLZINGER A, GEIERHOFER R, ERRATH M: *Semantische Informationsextraktion in medizinischen Informationssystemen.* Informatik-Spektrum 30(2) (pp. 69-78), Springer, 2007.

[2] DENECKE K, KOHLHOF I, BERNAUER J: *Use of multiaxial indexing for information extraction from medical texts.* In: Proc. Workshop on Foundations of Clinical Terminologies and Classifications; Romania, 2006.

[3] DENECKE K, KOHLHOF I: *Informationsextraktion aus medizinischen Texten basierend auf einer multiaxialen Indexierung.* Deutsche Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie e.V. (gmds). 51. gmds-Jahrestagung, Leipzig, 10.-14.09.2006.

[4] WINGERT F: Morphologic Analysis of Compound Words. Meth. Inform. Med. 24 (1985) 155-162. Stuttgart 1985.

[5] WINGERT F: Automated Indexing Based on SNOMED. Meth. Inform. Med. 24 (1985) 27-34. Stuttgart 1985.

[6] WINGERT F: *SNOMED. Systematisierte Nomenklatur der Medizin*. Hrsg. der amerikanischen Ausgabe R. A. Côté. Deutsche Ausgabe bearbeitet und adaptiert von F. Wingert. Berlin, Heidelberg, New York etc.: Springer. 1984.

[7] EFTHIMIADIS E: Query expansion. In WILLIAMS ME (Ed.), Annual review of information systems and technology (ARIST). Vol. v31. Information Today (pp. 121-187), 1996.

[8] ARONSON AR, RINDFLESCH TC, BROWN AC: Exploiting a Large Thesaurus for Information Retrieval. In Procedings RIAO 94, 1994.

[9] KINGSLAND LC, HARBOURT AM, SYED EJ, SCHUYLER PL. Coach: applying UMLS knowledge sources in an expert searcher environment. Bull Med Libr Assoc. 1993 April; 81(2): 178–183.

[10] VOLK M, RIPPLINGER R, VINTAR S, BUITELAAR P, RAILEANU D, SACALEANU B: Semantic Annotation for Concept-Based Cross-Language Medical Information Retrieval. *Intl. J. of Med. Informatics*, Volume 67:1-3, 2002

[11] FAULSTICH LC, MÜLLER F, SANDER A, PITZLER R, ERRATH M, HOLZINGER A: Semantisches Retrieval medizinischer Freitexte, erschienen in 53. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (gmds). Stuttgart, 15.-19.09.2008. Düsseldorf: German Medical Science GMS Publishing House; 2008. Doc MDOK1-4.