Extraktion von Lungenfunktionsparametern aus Arztbriefen

Martin Toepfer[1], David Schmidt[1], Georg Dietrich[1], Maximilian Ertl[1,2], Georg Fette[1,2], Mathias Kaspar[2], Stefan Störk[2], Frank Puppe[1]

[1] Universität Würzburg,
Institut für Informatik,
97074 Würzburg
{martin.toepfer,georg.dietrich,frank.puppe}@uni-wuerzburg.de,
david.schmidt@informatik.uni-wuerzburg.de

[2] Universität Würzburg,
Deutsches Zentrum für Herzinsuffizienz,
97078 Würzburg
{georg.fette}@uni-wuerzburg.de,
{ertl_m,stoerk_s,kaspar_m}@ukw.de

Einleitung und Fragestellung

Trotz fortschreitender digitaler Messwerterfassung liegen viele Informationen über Patienten nur in Form von Freitext vor. Diese Daten sind von hohem Interesse, u.a. zur Rekrutierung von Patienten für klinische Studien, Data Mining, oder Decision Support Systeme. Das zentrale Arztbriefarchiv des Universitätsklinikums Würzburg beispielsweise enthält über 1 Mio Arztbriefe (Zeitraum ab 1999) in denen über 40.000 Abschnitte zu Lungenfunktionstests (inklusive Spirometrie, Bodyplethysmographie, Blutgasanalyse) erkannt wurden. Die dort beschriebenen Parameter und Beurteilungen sind relevant für klinische Studien, aber ihre Erschließung durch Informationsextraktionsmethoden wurde bisher wenig untersucht. Sofern vorhanden, müssen Messwerte erkannt, andererseits aber auch Sprachkonstrukte interpretiert und auf vorher spezifizierte Begriffe abgebildet werden. Für den folgenden einfachen Textausschnitt

„Lungenfunktion: VC mit 2,22l = 33%, FEV1 mit 4,44l =55%. Beurteilung: keine Restriktion. Leichte zentrale, deutlich periphere Obstruktion." [Werte willkürlich gesetzt]

sollen beispielsweise die Attribut-Wert-Paare „Vitalkapazität [l]: 2,22", „Vitalkapazität [%]: 33", „Einsekundenkapazität [liter]: 4,44", „Einsekundenkapazität [%]: 55", „restriktive Ventilationsstörung: nein", „zentrale obstruktive Ventilationsstörung [vorhanden]: ja", „zentrale obstruktive Ventilationsstörung [Schweregrad]: leicht", „periphere obstruktive Ventilationsstörung [vorhanden]: ja", „periphere obstruktive Ventilationsstörung [Schweregrad]: ausgeprägt" ausgegeben werden. Als Voraussetzung müssen insbesondere Terminologien zur Informationsextraktion entwickelt werden, die die relevanten Terme, Konzepte und ihre Beziehungen spezifizieren. Darüber hinaus setzt die Verarbeitung von speziellen klinischen Domänen typischerweise die Anpassung von allgemeineren Segmentierungs-, Vor- und Nachverarbeitungskomponenten voraus.

Material und Methoden

Zur Entwicklung der Terminologie und der Informationsextraktionskomponente wurde ein speziell auf diesen Zweck ausgerichtetes Tool [1] benutzt und weiterentwickelt. Die Entwicklungsumgebung enthält vielfältige Funktionen, u.a. um Konzeptvorschläge aus Texten automatisch erstellen zu lassen, und anschließend die Konzepte, ihre Eigenschaften und Beziehungen zueinander, sowie die relevanten Ausdrücke zu verwalten. Weiterhin können über die Oberfläche Texte effizient und komfortabel durchsucht werden. Schließlich lassen sich Referenzannotationen setzen und damit Goldstandards zur Evaluation der konstruierten Komponente definieren. Der integrierte Informationsextraktionsalgorithmus benutzt die in der Terminologie angegebenen flachen

Beziehungen (Objekt-Attribut-Wert Strukturen und Templates) zwischen Konzepten sowie eine hierarchische Segmentierung, um mehrdeutige Ausdrücke korrekt zuzuweisen.

Für die Verarbeitung der Lungenfunktionstests wurden die für die Segmentierung zuständigen Regelskripte sowie die Vor- bzw. Nachverarbeitungsskripte für die Domäne angepasst. Unter anderem wurden Regeln zur Satzanalyse ergänzt, beispielsweise um Konstruktionen wie „Ventilationsstörung […] sowie verminderter PEF und MEF 50 mit 70 % bzw. 80 % der Norm“ korrekt zu behandeln. Hier überkreuzen sich Zuweisungen von Attributen und Werten (PEF=70%, MEF 50=80%).

Durch die Abschnittserkennung wurden aus dem Zeitraum ab 1999 insgesamt 41.468 (davon 40.762 allgemeine, und 706 als Spirometrie ausgezeichnete) Lungenfunktionsabschnitte gefunden, wobei ein Abschnitt jeweils mehrere relevante Unterabschnitte (z.B. Untersuchungen mit Datumsangabe) enthalten kann. Die Terminologie wurde zunächst auf aggregierten Phrasen aus einem Pool von 1.000 Abschnitten der Medizinischen Klinik 1 (ca. 17.000 Abschnitte) erstellt, und mit 69 nicht-aggregierten Dokumenten verfeinert. Anschließend wurden aus 1.000 weiteren Dokumenten (Datensatz *D.1*) 30 Dokumente *(D.1.30)* ausgewählt und zur Verbesserung der Unterabschnittserkennung benutzt. Die Abschnitte aus *D.1* und *D.1.30* dienten auch dazu Häufigkeiten von Layouteigenschaften abzuschätzen. Zum Testen wurden 50 weitere Abschnitte (nicht in *D.1* enthalten) aus der Gesamtmenge zufällig ausgewählt.

Zur Evaluation wurden true positives (TP; im Goldstandardsegment war die gefundene Attribut-Wert-ID angegeben), false positives (FP; gefundene ID war nicht angegeben) und false negatives (FN; Attribut-Wert-ID oder andere relevante Information aus dem Goldstandardsegment wurde nicht gefunden) gezählt. Falls pro Goldstandardsegment mehrere Annotationen mit der gleichen Attribut-Wert-Paar-ID auftraten, wurde es für die Evaluation nur einfach gezählt. Als Kennwerte wurden precision=TP/(TP+FP), recall=TP/(TP+FN), sowie der F1-Score=2*precision*recall/(precision+recall) berechnet; dies erfolgte einerseits für alle Attribut-Wert-Paar-Annotationen (micro-average), andererseits als Mittelwert über die F1-Scores bestimmter Konzepte (macro-average).

Ergebnisse

Aus den 30 Trainingsdokumenten (*D.1.30*) enthielten etwa die Hälfte (14 Dokumente) ausschließlich natürlichsprachliche Ausdrücke; von ihnen waren 7 kurze Normalbefunde. Ein Dokument enthielt im Fließtext beschriebene Messwerte. Auch im größeren Trainingsset *D.1* (1.000 Dokumente) wurden in knapp unter der Hälfte der Abschnitte (46%, 458 Dokumente) keine Schlüsselwörter von Messwerten erkannt.

Die entwickelte Terminologie enthält (ohne Auflösung von Templatereferenzen) 16 Templates, 21 Objekt-, 144 Attribut- und 254 Wertkonzepte, und insgesamt 561 verschiedene Varianten. Die Konzepte sind auf oberster Ebene aufgeteilt in „Beurteilung“ und „Messwerte“, wobei die wichtigsten Aspekte unter einem Terminologieknoten „Kernkonzepte“ gruppiert wurden.

Ohne weitere Anpassungen wurde auf den Trainingsdokumenten aus *D.1.30* ein F1-Score von 85% erreicht. Zu den Fehlerquellen zählten u.a. Segmentierungsfehler, falsch extrahierte Konzepte wenn ein Teilwort auftrat, fehlende Varianten, sowie Schreibfehler. Durch Anpassungen wurde der F1-Score auf diesem Trainingsset auf 97,3% verbessert. Auf dem Testset lag der micro-average F1-Score anschließend bei 93,1% (prec.=93,9%,rec.=92,3%). Beurteilungen zu Obstruktion/Restriktion

(vorhanden/nicht-vorhanden) wurden mit einem F1-Score (macro-average) von 96% erkannt, Messwertangaben zu FEV1 mit 99%.

Diskussion

Obwohl Lungenfunktionstests zumeist digital durchgeführt werden liegen die Ergebnisse oft ausschließlich textuell innerhalb von Arztbriefen vor. Die betreffenden Abschnitte enthalten zum Teil nur Messwerte oder aber auch allein natürlichsprachliche Ausdrücke. Gerade im letztgenannten Fall sind spezielle Informationsextraktionsanwendungen nötig, um die Information sinnvoll in Datenbanken zu überführen. Die aktuelle Arbeit ist eine Vorstudie, bietet jedoch bereits eine Terminologie und Regeln, mit denen sich die relevanten Standardparameter extrahieren und für anschließende Abfragen in das klinische Data Warehouse der Universität Würzburg [2] integrieren lassen. Temporale Aspekte wurden bisher nicht ausführlich betrachtet; sie stellen einen interessanten Punkt für nachfolgende Arbeiten dar, zum Beispiel die Behandlung von Vergleichswerten. Darüber hinaus können unter anderem Parser-Technologien, sowie die Integration und der Vergleich von Informationen aus unterschiedlichen Abschnitten des Arztbriefes weiter untersucht werden.

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

Literatur

[1] Toepfer M, Fette G, Beck PD, Klügl P, Puppe F. Integrated Tools for Query-driven Development of Light-weight Ontologies and Information Extraction Components. In: Ide N, Grivolla J, editors. Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for HLT; 2014 Aug 23; Dublin, Ireland. Association for Computational Linguistics and Dublin City University; 2014. p. 83-92.

[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. In: Goltz U, Magnor M, Appelrath HJ, Matthies HK, Balke WT, Wolf L, editors. Lecture Notes in Informatics 208: Proceedings of Informatik 2012; 2012 Sep 16-21; Braunschweig, Germany; Bonn: Gesellschaft für Informatik; 2012. p. 1237-1251.