

Das IT- und Medienstipendium für innovative Projekte von Studierenden

Philipp Daumke

Projekt: Multilinguale medizinische Suchmaschine für das WWW

Projektbeschreibung:

Entwickelt wurde ein multilinguales Suchverfahren für biomedizinische Dokumente im WWW. Es bietet Lösungsansätze für zahlreiche Herausforderungen der medizinischen Textrecherche wie Umgang mit (Mehrwort-) Synonymen und Kompositabildung. Das Verfahren basiert auf dem Morphosaurus-System, welches Texte mittels eines mehrsprachigen Lexikons in morphologisch sinnvolle Einheiten (Subwörter) zerlegt. Die Übersetzung der Anfrage erfolgt durch eine Transformation in diese Subwörter und eine anschließende Übersetzung in die Zielsprache mittels großer monolingualer Wortlisten. Diese Übersetzung wird an eine Standardsuchmaschine gesendet und die Ergebnisse dem Benutzer über ein Webinterface präsentiert.

Hintergrund:

Ich bin derzeit im Doktorandenstudium in der Abteilung für medizinische Informatik am Universitätsklinikum. Der Forschungsschwerpunkt unserer Arbeitsgruppe liegt im Bereich Information Retrieval. Besonderes Interesse wird auf die mehrsprachige Suche (engl. *Cross Language Information Retrieval, CLIR*) insbesondere in der biomedizinischen Domäne gelegt. In den letzten Jahren wurde erfolgreich das MorphoSaurus-System implementiert, welches auf begrenzten Dokumentensammlungen

eine effiziente Dokumentenrecherche ermöglicht.

Ziel meines Projektes war, basierend auf diesem System ein Recherchesystem für das WWW zu implementieren, welches sich auf die mehrsprachige Dokumentenrecherche im Bereich der Biomedizin konzentriert. Die riesige und inhomogene Dokumentensammlung des WWW erfordert hierbei völlig neue Ansätze, um eine entsprechende Performanz bei der Recherche zu gewährleisten.

Ziele:

Zu den konkreten Zielen des Projektes zählt eine hochwertige Übersetzung von

Benutzeranfragen in andere Sprachen. Dies soll mithilfe großer medizinischer Textkollektionen aus verschiedenen Online-Quellen ermöglicht werden, die mithilfe des Morphosaurus-Systems sprachlich verknüpft werden. Ein zweiter Schwerpunkt liegt in der Entwicklung einer schnellen Datenbankbindung, welche Anfragen innerhalb von Sekunden bearbeitet. Schließlich soll das System wissenschaftlich evaluiert und die Ergebnisse veröffentlicht werden.

Realisierung der Suchmaschine:

Zunächst wurden aus verschiedenen frei verfügbaren, domänen- und sprachspezifischen Quellen große biomedizinische Textcorpora zusammengestellt. Diese berücksichtigen in abnehmender Corpusgröße die Sprachen Englisch, Deutsch, Spanisch, Portugiesisch und Schwedisch. Aus den Corpora werden durch einen Tokenizer Listen von Einzelwörtern sowie von Wortpaaren und -tripeln benachbarter Wörter gebildet (sog. Übersetzungseinheiten). Als zusätzliche Information enthalten diese Listen die Häufigkeiten des Auftretens der Übersetzungseinheiten in den Textcorpora. Mithilfe des Morphosaurus-Systems werden die Einheiten anschließend sprachlich verknüpft, so dass Gruppen von Übersetzungen mit gleicher sprachlicher Bedeutung entstehen. Die Datenbank enthält mehrere Millionen solcher Einträge.

Ein Benutzer sendet über ein Web-Interface Anfragen an das Recherche-System. Diese Anfrage wird nun mit den Einträgen in der Datenbank verglichen und diejenigen Einträge herausgefiltert, die der Anfrage in der vom Benutzer gewünschten Sprache am ehesten

entsprechen. Diese werden anschließend an eine Standard-WebSuchmaschine geschickt. Abbildung 1 gibt einen Überblick über den Datenfluss innerhalb des Systems.

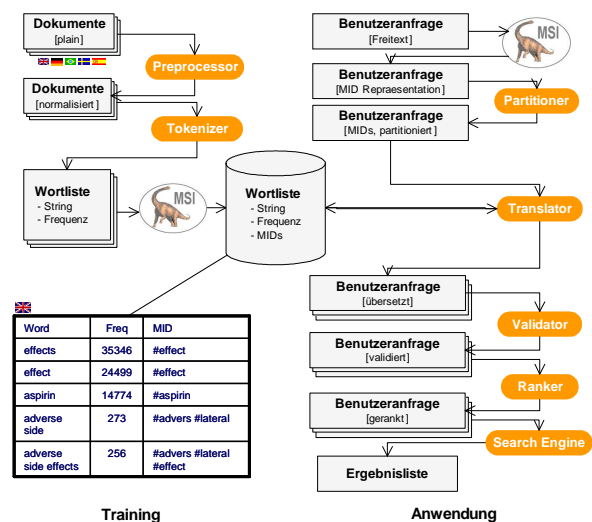


Abbildung 1: Übersicht über den Datenfluss in der Suchmaschine

Die technische Realisierung des Frontends und der Übersetzungslogik erfolgte mit HTML und Perl. Als Datenbank wurde die Berkeley-DB verwendet, die eine hochperformante Suche auf den Übersetzungseinheiten ermöglichte. Abbildung 2 ist ein Screenshot der Weboberfläche unserer Suchmaschine.

Lexikalische Ressourcen:

Das Morphosaurus-System, welches die Textlisten in eine sprachunabhängige Darstellung überführt, beinhaltet als lexikalische Ressourcen sogenannte Subwort-Lexika. Diese werden manuell erstellt und gepflegt und wurden mit Hilfe dieses Projektes weiter verbessert. Lexikographen am Universitätsklinikum Freiburg kontrollierten insgesamt ca. 5000 vom System generierte Übersetzungen und passte die lexikalischen Ressourcen

bei falschen Übersetzung entsprechend an.

Ergebnisse:

Das Verfahren wurde auf zwei medizinischen Kollektionen getestet, die speziell für die Evaluation medizinischer Recherchesysteme konzipiert wurden (siehe Tabelle 1).

Bei diesen Ergebnissen besonders erfreulich ist, daß das im Projekt entwickelte Verfahren in der mehrsprachigen Suche (bei deutschen Anfragen auf englischen Dokumenten) eine durchschnittliche Präzision erreicht, die der monolingualen Baseline ebenbürtig (Ohsumed) bzw. sogar überlegen (ImageCLEFmed) ist.

Die Werte geben den Stand der Arbeiten an unseren lexikalischen Ressourcen wieder. Diese sind für das Deutsche und Portugiesische recht weit und müssen insbesondere für die anderen Sprachen weiter fortgesetzt werden.

Sprache	Ohsumed – 11pt Precision	ImageCLEFmed – 11pt Precision
Baseline	.17	.16
Englisch	.19 (112%)	.15 (94%)
Deutsch	.17 (100%)	.17 (106%)
Portugiesisch	.15 (88%)	.16 (100%)
Spanisch	.09 (53%)	.07 (44%)
Schwedisch	.11 (65%)	.12 (82%)

Tabelle 1: 11pt Precision-Ergebnisse über zwei Testkollektionen



Abbildung 2: Screenshot der Weboberfläche der Suchmaschine

Veröffentlichungen

Die Ergebnisse des Projektes wurden auf einigen Konferenzen und Tagungen präsentiert, unter anderem auf dem 10. Weltkongress *Internet in Medicine* in Prag sowie beim Mannheimer Forschungstag *doIT Software-Forschungstag 2006* in Mannheim. Die folgende Liste gibt einen Überblick über die mit diesem Projekt verbundenen Publikationen:

Daumke P., Schulz S., Markó K: Morphoogle - Eine medizinische CLIR Schnittstelle zu einer Web-Suchmaschine. In: R. Klar, W. Köpcke, K. Kuhn, H. Lax, A. Zaiß (eds.): Tagungsband der 50. Jahrestagung der Deutschen Gesellschaft für Medizinische Informatik, Biometrie und Epidemiologie (GMDS '05), Freiburg, Germany. 2005: 246-248

Daumke P, Schulz S, Markó K: A CLIR Interface to a Web Search Engine, Proceedings of AMIA Symposium 2005, Washington D.C., 2005: 934

Daumke P, Schulz S, Markó K: Searching Multilingual Medical Content in the Web. *Technology and Health Care*, 2005; 13 (5): 375-376

Daumke, P., Schulz S & Markó K (2006). Subword approach for acquiring and cross-linking multilingual specialized lexicons. In LREC 2006 – Proceedings of the 5th International Conference on Language Resources and Evaluation

Workshop: Acquiring and representing multilingual, specialized lexicons: the case of biomedicine.

Daumke P., Markó K, Schulz S & Poprat M. (2006). Morphoogle - Eine multilinguale medizinische Suchmaschine für das WWW. In Heinzl A. Klumpp D. Haasis, K. (Ed.), Aktuelle Trends in der Softwareforschung - Tagungsband zum doIT Software-Forschungstag 2006, pp. 133–142.

Zusammenfassung

Dieses Projekt bietet einen vielversprechenden Ansatz zur sprachübergreifenden Suche nach biomedizinischen Dokumenten im WWW. Erreicht wurde ein schnelles Verfahren der Übersetzung von Benutzeranfragen und eine Präzision bei der Übersetzung von bis zu 106% in der mehrsprachigen Evaluation. Zur weiteren qualitativen Verbesserung der Ergebnisse werden derzeit Verfahren der natürlichen Sprachverarbeitung wie Wortartenerkennung getestet.

Mithilfe des MFG-Stipendiums war eine Realisierung dieses Projektes in dem oben beschriebenen Umfang möglich. Es bot sich insbesondere die Gelegenheit, das Projekt auf einigen internationalen Konferenzen zu präsentieren. Ich danke der MFG sehr herzlich für die große Unterstützung.

Karl-Steinbuch-Stipendium

MFG Stiftung

Karl-Steinbuch-Stipendium

Breitscheidstr. 4

70174 Stuttgart

Tel. +49/711/90715/314

stiftung@mfg.de

Über das Stipendiumprogramm

Mit dem Karl-Steinbuch-Stipendium fördert die MFG Stiftung

Baden-Württemberg innovative wissenschaftliche und

künstlerische IT- und Medienprojekte, die Studierende aus

Baden-Württemberg zusätzlich zu Ihrem Studium durchführen.

Die Projekte dauern 6-12 Monate und werden mit bis zu

9.600 € gefördert.

Weitere Informationen:

www.karl-steinbuch-stipendium.de