Mai / Juni 2012 Ausgabe #2-12 Deutschland Euro 12,00 ISSN: 1864-8398 www.dokmagazin.de

Verlängert! Jubiläumsabo

DOK.

Technologien, Strategien & Services für das digitale Dokument

Digital Publishing

Die neue Welt des Publizierens

O/M - Intelligente Dokumente sind gefragt:
E-Presentment für alle Ausgabekanäle

Special **Wissensmanagement**
Satzsemantische Suche / Personalisierte Inhalte

Praxis Von der E-Mail zu Social Business-Lösungen

Satzsemantische Suche – präziser finden mit der TAKE Searchbench

Enterprise Search, Volltext- und Metadatensuche, Wortartanalyse, Satzanalyse

www.dfki.de/~uschaefer

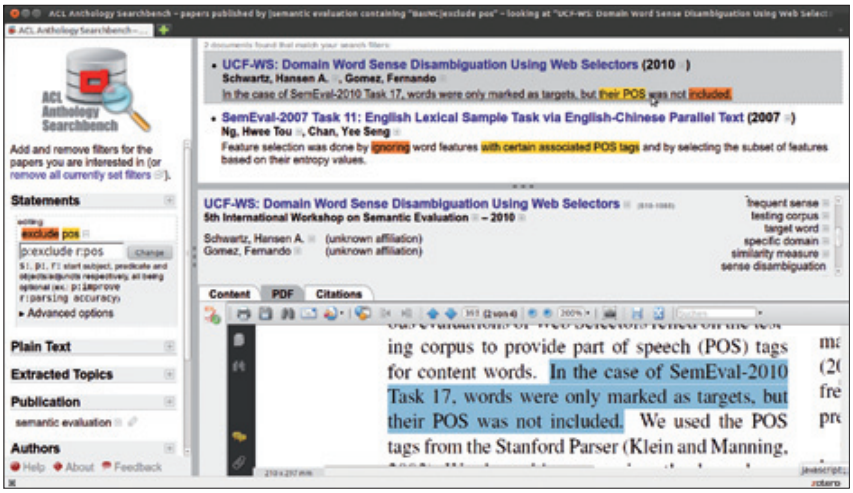
Dr. Ulrich Schäfer ist Senior Engineer und Projektleiter am **Language Technology Lab im Deutschen Forschungszentrum für Künstliche Intelligenz GmbH** (DFKI) in Saarbrücken. Seit 2000 forscht er an hybriden Sprachtechnologieverfahren für multilinguale Informationsextraktion aus Texten, Automatische Fragebeantwortung und Semantische Suche. Davor war er fast fünf Jahre Anwendungsentwickler und Consultant für electronic messaging, EDIFACT und multilinguale Office Automation Software vor allem für EU-Institutionen und Industriekunden in Luxemburg, Belgien und Deutschland tätig.

Mit zunehmenden Dokumentmengen wird präzisionsorientierte Suche immer wichtiger. Die berühmte Stecknadel im Heuhaufen lässt sich allein mit Volltextsuche – im übertragenen Sinne – kaum mehr finden. Im BMBF-geförderten Projekt TAKE ("Technologies for Advanced Knowledge Extraction") am Deutschen Forschungszentrum für Künstliche Intelligenz in Saarbrücken wird daher an semantischen Textanalyseverfahren geforscht, mit denen unter anderem eine zielorientierte Suche in Dokumenten möglich wird. Die Textanalyse ermöglicht auch zahlreiche weitere Anwendungen wie automatische Informationsextraktion, Taxonomie- und Ontologieextraktion sowie Fragebeantwortung.

Die TAKE Searchbench ist eine Suchanwendung, die beispielhaft 22.500 PDF-Dokumente, davon ein Drittel gescannte, inhaltlich suchbar macht. Sie ist frei zugänglich unter http://take.dfki.de/#Systems.

Nach Aussagen suchen

Die Idee der sogenannten satzsemantischen Suche besteht darin, zusätzlich zur Metadaten- und Volltextsuche auch nach Aussagen suchen zu können. Das bedeutet, dass Treffer nur dann angezeigt werden, wenn die Worte der Anfrage im Sinnzusammenhang in einem Satz vorkommen, nicht aber, wenn sie nur zufällig nahe beieinander liegen. Die Suchanfrage wird dazu in eine Subjekt-Prädikat-Objektstruktur gestellt, also z.B. **s**:semantics **p**:helps **r**:retrieval.



Suchinterface – mit dem Adobe Acrobat Reader Browser-Plug-in werden Suchergebnisse sogar im Originallayout hervorgehoben angezeigt

Dabei stehen

s: für Subjekt,

p: für Prädikat,

r: für "Rest"

(verschiedene semantische Objekte, die der Benutzer nicht im Detail angeben muss, z.B. direkte, indirekte Objekte sowie Adjunkte).

Gefunden wird mit der oben genannten Beispiel-Query beispielsweise ein Satz wie *"More than this, Schutze and Pedersen (1995) performed experiments which have shown that semantics can actually help retrieval performance."*

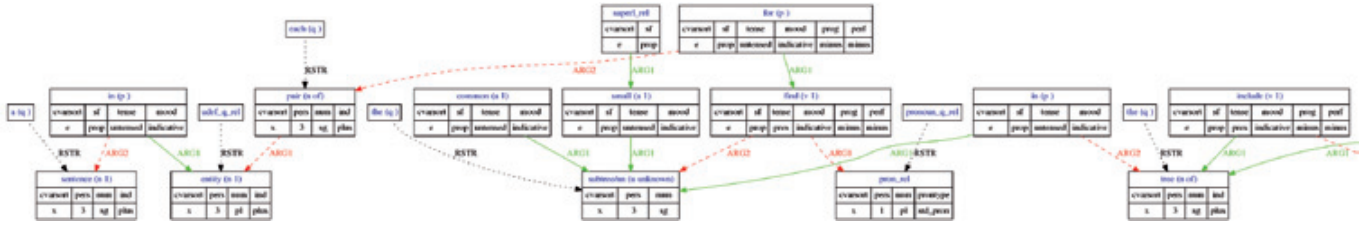
Satzsemantisch zu suchen bedeutet, dass es auf den Sinn ankommt, nicht auf die syntaktische Struktur. In der Searchbench werden daher Aussagen, die im Passiv ausgedrückt sind, mit ihren aktiv formulierten Entsprechungen gleichgesetzt. So findet man beispielsweise unter der Abfrage **p**:improve **r**:performance auch Aussagen wie *„Performance was improved by 34%"*.

Wie man an diesem Beispiel sieht, dürfen Teile der Fragestruktur in der Query weggelassen werden. Man kann also demzufolge nach Subjekt-Objekt-Strukturen suchen, ohne das Prädikat zu spezifizieren. Bei der Eingabe von **s**:Peter **r**:Mary wird man in den Texten daher alles finden, was Peter und Mary Verschiedenes miteinander tun. Und umgekehrt: Da die häufigste Anfragestruktur von der Form „Prädikat und Rest" (ohne Spezifikation des Subjekts) ist, darf man diese ebenfalls genau so eingeben: Die Anfrage helps text retrieval ist also identisch mit **p**:helps **r**:text retrieval.

Noch präziser suchen

Zusätzlich ist es möglich, negierte Aussagen auszublenden, um die Anzahl der Treffer zu verringern. Andererseits kann durch den Einbezug von Synonymen die Trefferquote erhöht werden. Übrigens: Beides zusammen ist die Voreinstellung bei der Suche mit der TAKE Searchbench. Darüber hinaus lassen sich negierte Antonyme als Synonyme finden. Ein Beispiel: Für die Suchanfrage exclude POS wird auch *„POS was not included"* gefunden (siehe Bild 1)! Als Quelle für die Synonyme und Antonyme wird gegenwärtig WordNet verwendet, eingeschränkt auf die statistisch häufigsten Bedeutungen in der indizierten Textbasis.

Da die Suchergebnisse sich auf die in der Query geforderte semantische Position beschränken, werden im Allgemeinen weniger Treffer angezeigt als bei der Volltextsuche mit den gleichen Worten. Dieser Effekt ist in der Regel gewünscht, denn die Präzision der gefundenen und angezeigten Inhalte steht bei TAKE Searchbench im Vordergrund.



Beispiel für die „feinkörnige“ semantische Repräsentation

Verbindung mit klassischer Volltext- und Metadatensuche

Ob viele oder wenige Treffer erzielt werden, hängt unter anderem davon ab, wie spezifisch die Anfrage formuliert wurde. Daher ist es manchmal sinnvoll, die semantische Suche mit anderen, klassischen Suchverfahren wie Metadaten (Verfasser, Titel, usw.) mit autosuggest-Eingabefeldern sowie Volltextsuche zu verbinden. All dies ist in der Searchbench realisiert – daher ihr Name: eine “Werkbank” für die Suche.

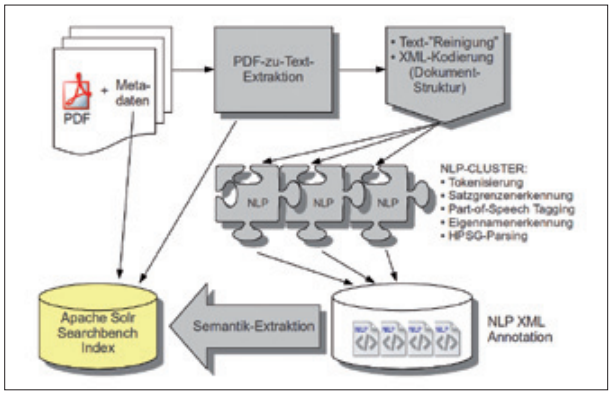
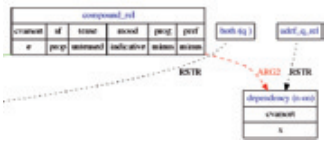


Diagramm Offline-Analyseprozess

Systemarchitektur

Wie wird nun so ein satzsemantischer Index erzeugt? Im Prinzip handelt es sich um einen Offline-Prozess, der einen Apache-Solr-Index mit einer speziellen Struktur für die effiziente Speicherung der Satzobjekte und ihrer assoziierten Texte füllt (siehe Bild 2). Dazu wird zuerst der Text aus den PDF-Dokumenten – andere Dokumentformate sind natürlich möglich – in ein XML-Format überführt. In einer parallelisierten Vorverarbeitungs-Pipeline wird nun der Text tokenisiert und in Sätze zerlegt. Unter anderem wird hier zur Volltextindizierung abgezweigt (Pfeil) und die Position des Satzes einschließlich Seitenzahl gemerkt. Dies ermöglicht es, den Satz als Suchergebnis später im Acrobat Reader anzuzeigen und zu markieren.



Zusätzlich wird eine statistische Wortartanalyse (“POS-Tagging”) vorgenommen. Diese ist wichtig für die Robustheit bezüglich unbekannter Worte. Denn wird ein Wort nicht im Lexikon des semantischen Parsers gefunden, kann so seine wahrscheinliche Wortklasse und damit seine ungefähre “Bedeutung” im Satzzusammenhang geraten werden. Auf diese Weise wird das System auch robust gegenüber Schreibfehlern. Gemeinsam mit den Resultaten einer Eigennamen-Erkennungskomponente werden diese Informationen an den HPSG-Parser übergeben, der jeden Satz syntaktisch analysiert und gleichzeitig eine feinkörnige semantische Repräsentation erzeugt (siehe Bild 3). Von dieser wird dann die für die Indizierung stark vereinfachte Subjekt-Prädikat-Objekte-Struktur abgeleitet und an Solr übergeben.

Das Suchinterface der Searchbench ist eine auf GWT und weiteren Open Source JavaScript-Frameworks basierende Browseranwendung, die mit dem in Java und rund um Apache Solr realisierten Server über HTTP kommuniziert.

Fazit

Mit computerlinguistischen Verfahren zur Satzanalyse, die weit über die bekannten Möglichkeiten der Volltextsuche hinausgehen, können Suchindizes für eine präzise semantische Suche erstellt werden. In Apache Solr verwaltet, lassen sich damit neuartige Suchanwendungen und weitere semantikorientierte Anwendungen wie automatische Fragebeantwortung erstellen. Geeignet ist die weitgehend domänenunabhängige Technologie für alle redigierten, grammatikalisch überwiegend wohlgeformten Texte, beispielsweise in digitalen Bibliotheken, Nachrichtenarchiven, Patenttexten, technischen Dokumenten, wissenschaftlichen Veröffentlichungen.

Ausblick

In einer der nächsten Ausgaben soll über weitere interessante Anwendungen berichtet werden, die durch Kombination von rein statistischen Verfahren der Termextraktion mit satzsemantischer Analyse möglich sind: vollautomatische Glossar-Erstellung sowie die Extraktion semantischer Netzwerke wie Taxonomien und Ontologien über inhaltliche Bereiche (Domänen) aus Texten – ohne vorherige Wissensmodellierung der jeweiligen Domäne. In einem weiteren Artikel stellen wir eine grafische Suche vor: Dokumentabhängigkeiten, beispielsweise Zitationen, wie sie in wissenschaftlichen, juristischen, aber auch in technischen Dokumenten vorkommen, können aus den Texten bzw. Metadaten extrahiert werden – die daraus resultierenden Graphen werden zur Darstellung und Navigation verwendet.

Informare!

Informare!-Keynote von **Hans Uszkoreit**, Bereichsleiter Sprachtechnologie im DFKI am 8. Mai 2012, 10.30 Uhr: *“Turings Traum weiter träumen: Mit Sprachtechnologie und KI auf dem Weg zur Social Intelligence”*. Mehr Infos auf

www.informare-wissen-und-koennen.com.