Januar / Februar 2013 Ausgabe #6-12 Deutschland Euro 12,00 ISSN: 1864-8398 www.dokmagazin.de

DOK.

Technologien, Strategien & Services für das digitale Dokument

Archivierung – sicher & klar strukturiert

Big Data, E-Mail- und Web-Archivierung

Semantische Systeme für Markt- & Trendanalyse

Special **E-Invoicing**
Step-by-Step zur E-Invoicing Plattform

ECM mit SharePoint 2013 – smart & social

Interview Managed Print Services: „Es führt kein Weg daran vorbei."

Automatische Terminologie-, Taxonomie- und Glossarextraktion

Begriffshierarchien, Schlüsselwörter, Klassifizierung von Dokumenten, Extraktionsmuster, Redundanz

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, EDI und multilingual office automation tätig.

Zum Thema „Suche" wurde an dieser Stelle bereits eine Technologie vorgestellt, die auf satzsemantischer Analyse von Textdokumenten basiert (DOK.magazin 2/2012): In der TAKE Searchbench wird nach ähnlichen Aussagen und in strukturierter Form (semantisches Subjekt, Prädikat und Objekt) gesucht, so dass Ergebnisse präziser angezeigt werden können. Ferner lässt sich das Ganze mit einer Volltext- und Metadatensuche kombinieren. Eine weitere Möglichkeit besteht darin, nach Schlüsselworten (Themen) zu fragen, die mittels statistischer Verfahren pro Dokument extrahiert werden. Der folgende Artikel wird sich zunächst um das Verfahren für diese Extraktion drehen, sowie anschließend um die darauf basierenden weiteren interessanten Anwendungen wie die automatische Erstellung von Glossaren und Taxonomien (Begriffshierarchien).

Beginnen wir mit einem Beispiel für die Terminologieextraktion: In einem Artikel über Automotoren könnten dazu Begriffe wie „Benzinmotor", „Dieselmotor", „Ottomotor" oder „Wankelmotor" automatisch als wichtig identifiziert werden. Ferner sollen „Ottomotor" und „Wankelmotor" als Unterbegriffe von „Benzinmotor" und dieser wiederum als „Motor" identifiziert werden. Möglich ist dies durch Erkennen von Definitionssätzen (z.B. „Ein Wankelmotor ist ein Benzinmotor mit einem Rotationskolben."). Das Resultat wäre dann ein kleiner Ausschnitt einer Domänentaxonomie. Ferner können im Rahmen der Glossarextraktion Sätze im Text automatisch gefunden werden, welche Definitionen, nähere Angaben oder Beurteilungen der vorgenannten Begriffe beinhalten, z.B. ein Satz der Art „Der Ottomotor wurde zu Ehren von Nicolaus August Otto benannt".

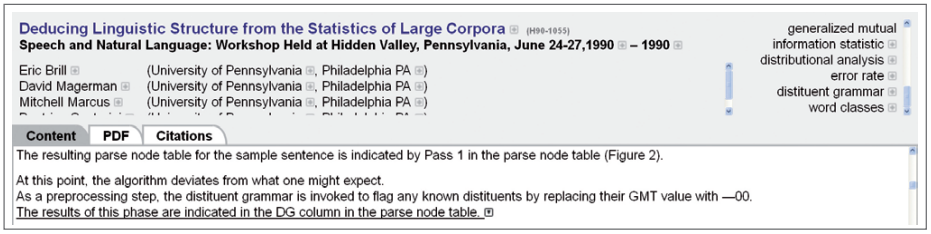

Bild 1: Je Dokument automatisch extrahierte Fachbegriffe in der TAKE Searchbench (rechts oben); (Quelle: http://take.dfki.de)

Basistechnologie: automatische Terminologieextraktion

Gemeinsame Grundlage für die automatische Glossar- und Taxonomieerstellung ist zunächst die Bestimmung von thematisch relevanten Begriffen aus der Inhaltsdomäne. Diese werden nicht vorgegeben, sondern aufgrund statistischer und linguistischer Eigenschaften aus dem Text berechnet. Als sehr robust hat sich hierbei die sogenannte C-/NC-Wert-Methode (Frantzi, Ananiadou und Mima, 2000) erwiesen; es gibt jedoch auch andere, ähnlich wirksame Verfahren. Die Eingabe für die sogenannte Terminologieextraktion besteht lediglich aus dem Text selbst. Die daraus automatisch extrahierten linguistischen Informationen (domänenunabhängig über die Wortarten im Text) und statistischen Eigenschaften geben anschließend Auskunft über Häufigkeiten und das Vorhandensein von Mehrwortgruppen im Text.

Wichtig und vorteilhaft bei diesem Verfahren ist, dass dabei neben der Ausgabe der Begriffe auch gleich eine Bewertung in Form eines relativen numerischen Wertes vorgenommen wird, so dass die wichtigsten Suchergebnisse zuerst ausgegeben werden– sie werden in der TAKE Searchbench pro Dokument berechnet und angezeigt (rechts in Bild 1, der Dokumentansicht). Um gleichzeitig möglichst viele unwahrscheinliche „Einzelgänger“ auszuschließen, wurde mit demselben Verfahren eine globale Liste auf Basis aller Texte berechnet. Damit konnten die jeweiligen Einzeldokument-Terme identifiziert werden. Grundsätzlich ist dieses Verfahren natürlich nur sinnvoll, wenn angenommen werden kann, dass die Texte dem gleichen thematischen Feld angehören. Eine manuelle Korrektur wurde in der Searchbench – schon aufgrund der großen Anzahl der Dokumente – nicht vorgenommen.

www.coli.uni-sb.de/~magda

Magdalena Wolska, MPhil, ist seit 2003 wissenschaftliche Mitarbeiterin am **Institut für Computerlinguistik der Universität des Saarlandes** in Saarbrücken. Sie forscht an der semantischen Sprachanalyse von technischer und wissenschaftlicher Sprache im Kontext von wissenschaftlichen Veröffentlichungen und intelligenten tutoriellen Dialog-Systemen. Davor war sie über drei Jahre lang Mitarbeiterin bei der Sprachverarbeitungsgruppe des Educational Testing Service in Princeton, NJ.

Bild 2: Invaders als Game with A Purpose (GWAP)

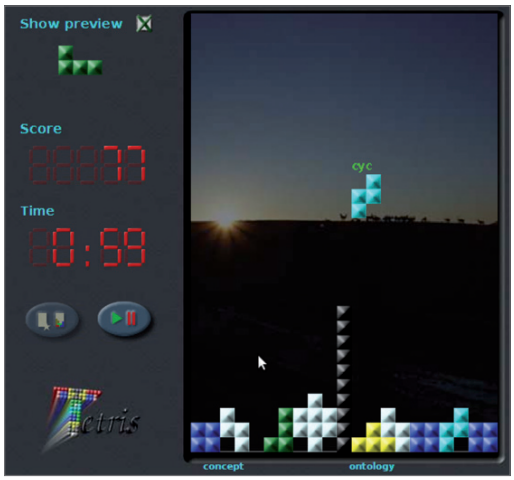

Bild 3: Tetris als Game with A Purpose (GWAP)

Taxonomieextraktion

Liegen genügend Dokumente zu einer Inhaltsdomäne elektronisch vor, so lassen sich aus den Texten auch Taxonomien, also Begriffshierarchien, ableiten. Diese wiederum sind eine wichtige Wissensquelle für die Strukturierung (Klassifizierung) von Dokumenten, für Suchaufgaben (z.B. ähnliche/speziellere/allgemeinere Begriffe finden), aber auch für die automatische Fragebeantwortung aus Texten. Der Idee liegt die Beobachtung zugrunde, dass z.B. technische oder wissenschaftliche Dokumente, aber auch allgemeinere Texte wie Nachrichten oft Begriffsdefinitionen enthalten.

Da wir die Begriffe unserer Domäne mittels Terminologieextraktion schon automatisch gesammelt haben, können wir leicht Sätze finden, die einen Begriff durch Verweise auf einen Ober- oder Unterbegriff definieren, z.B. „a car is a vehicle with four wheels“ oder „vehicles such as cars, motorbikes and bicycles“. Aus dem zweiten Satz lassen sich gleich drei is-a-Beziehungen ableiten. Außerdem wird die aus dem ersten Satz extrahierte Definition „car is a vehicle“ bestätigt. Jedoch werden bei diesem Verfahren auch falsche Beziehungen extrahiert – trotz ausgefeilter Extraktionsmuster und unter Ausnutzung von Redundanz (Verwendung nur solcher Paare, die mehrfach „definiert“ wurden). Dies liegt an der unendlichen Ausdrucksfähigkeit und Mehrdeutigkeit der natürlichen Sprache.

Wie können wir also die Qualität der automatisch extrahierten Definitionen sicherstellen? Nur dann macht die Anordnung der Paare in einer Hierarchie (Taxonomie) Sinn. Da wir Fachbegriffe extrahieren, sind für die Kontrolle zwingend Domänenexperten vonnöten. Unsere Erfahrung mit diesen Evaluationsaufgaben zeigte jedoch, dass oft nur wenige Experten bereit sind zu helfen. Diese sind in der Regel chronisch überlastet und haben weder die Zeit noch die Motivation, eine derart umfangreiche Wissenskontrolle auszuüben – insbesondere dann, wenn eine Maschine manchmal sinnlose Paare zugeordnet hat. Spielerisch sind sie aber doch zu gewinnen! Wir haben ein Experiment durchgeführt, das eindrucksvoll die Möglichkeiten sogenannter „serious games“ oder „games with a purpose“ (GWAP) für die Expertengewinnung demonstriert. Man spricht in diesem Zusammenhang auch von „crowdsourcing“, also der Hinzuziehung von Freiwilligen zur Lösung von verteilbaren Aufgaben.

Wir haben dazu zwei populäre browserbasierte Online-Spiele, Tetris und Alien Invaders, in Open-Source-Varianten so modifiziert, dass Experten sie zum Korrigieren bzw. Evaluieren der automatisch extrahierten „is-a“-Paare spielen konnten. Bei Invaders wurde pro Runde ein (automatisch extrahierter) Begriff vorgegeben; extrahierte Unterbegriffe kamen von oben als „Aliens“ wie in Bild 2 gezeigt. Falsche Unterbegriffe sollten abgeschossen werden, die richtigen überleben. Ähnlich bei Tetris, das in zwei Hälften geteilt, jeweils einen von oben fallenden Begriff auffangen sollte. Dieses Spiel war bei den Freiwilligen überraschenderweise deutlich beliebter als Aliens, aber weit vor einer dritten Spielvariante, einem einfachen Multiple-Choice-Quiz mit Checkboxen und Begriffslisten in einer simplen textbasierten Weboberfläche.

Es spielten etwa 60 Freiwillige innerhalb von zehn Tagen; zehnmal so viele wie bei einer vergleichbaren Evaluationsaufgabe mit nur einem „langweiligen“, textbasierten Benutzer-Interface. Die Begriffspaare wurden aus „Pools“ zufällig ausgewählt. Die Gruppierung der Paare in diesen Pools stellte darüber hinaus sicher, dass genügend Paare von unterschiedlichen Spielern bewertet

werden: Von etwa 10.000 automatisch extrahierten Begriffspaaren wurden im Spielzeitraum fast 3.000 bewertet. 77 Prozent der dreifach und 80 Prozent der fünffach bewerteten Paare wurden dabei von jeweils drei bzw. fünf Spielern gleich bewertet (Wolska et al, 2011). Die Begriffspaare, die dieser (mehrfachen) Überprüfung standhielten, eignen sich folglich als zuverlässige Taxonomie-Bausteine.

Automatische Glossarerstellung

Nach Wikipedia ist ein Glossar „eine Liste von Wörtern mit beigefügten Erklärungen oder Übersetzungen". Wie bei der Taxonomie- wird auch bei der automatischen Glossarerstellung als Basis eine Sammlung von Texten zum gleichen Thema eingesetzt. Das Ziel des Verfahrens ist es, möglichst gute Glossarsätze aus den Dokumenten zu extrahieren, die sinnvolle Erklärungen zu den Begriffen liefern. In der Regel sind dies mehrere Sätze; diese können auch eine Bewertung des Begriffs oder Beispiele enthalten.

Wir haben dazu zwei Verfahren entwickelt und auf derselben Dokumentmenge (knapp 19.000 wissenschaftliche Veröffentlichungen) evaluiert. Das erste basiert auf ähnlichen lexikalisch-syntaktischen Mustern wie die Taxonomieextraktion, enthält aber einige allgemeinere Muster und erweitert automatisch die Muster- und Instanzenmenge („bootstrapping"). Das zweite Verfahren – und hier schließt sich nun der Kreis zur eingangs erwähnten Searchbench – verwendet deren tiefe satzsemantische Analyse mit nur einem einzigen Muster: S(=Subjekt):Begriff, P(=Prädikat):is. Hier kommt die Eleganz des semantischen Index zum Tragen: Sätze, in denen „is" ausschließlich als Hilfsverb verwendet wird, werden damit als Suchergebnis ausgeschlossen. Die automatische tiefe Satzanalyse abstrahiert von solchen „Oberflächlichkeiten" und liefert nur Sätze, in denen das Verb „sein" als Hauptverb gebraucht wird, also „sein" im Sinne einer Definition oder Beurteilung. Die Evaluation mit Experten (Reiplinger et al., 2012) zeigt interessanterweise, dass beide Verfahren etwa gleich gut sind, obwohl das semantische ja nur eine einzige Art von Sätzen erfasst – diese aber mit hoher Präzision. Beide Verfahren sollten daher am besten kombiniert werden zu einem Bootstrapping-Verfahren, das tiefe satzsemantische Analyse verwendet und mit mehr Verben und Satzmustern startet als nur s:Begriff, p:is. Der große Vorteil der beschriebenen Verfahren liegt darin, dass sie automatisch und domänenunabhängig funktionieren.

Literatur (online unter http://take.dfki.de)
M. Reiplinger, U. Schäfer, M. Wolska:
Extracting Glossary Sentences from Scholarly Articles:
A Comparative Evaluation of Pattern Bootstrapping and Deep Analysis. Proceedings of the ACL-2012 Main Conference Workshop on Rediscovering 50 Years of Discoveries, pages 55-65. Jeju Island, Republic of Korea, 2012.
M. Wolska, U. Schäfer, T. N. Pham:
Bootstrapping a Domain-specific Terminological Taxonomy from Scientific Text. 9th International Conference on Terminology and Artificial Intelligence (TIA), pages 17-23, Paris, France, 2011.
K. Frantzi, S. Ananiadou, and H. Mima. 2000.
Automatic recognition of multi-word terms: the C-value/NC-value method. International Journal on Digital Libraries, 3:115–130.

Ein gehaltvoller Jahrgang 2012.

Jetzt Abo für 2013 sichern!

Print – E-Paper – App

DOK.

Technologien, Strategien & Services für das digitale Dokument

www.dokmagazin.de/abo