Automatische Erstellung eines Wörterbuchs für Abkürzungen

In diesem Projekt soll ein System zur automatischen Erkennung von Abkürzungen mit den zugehörigen Begriffen entwickelt werden. Die Abkürzungen werden in Texten in natürlicher Sprache gesucht und dem abgekürzten Begriff zugeordnet. Die Texte stammen aus einem homogenen Themengebietgebiet (voraussichtlich Informatik-Veröffentlichungen)

Die Effektivität des Verfahrens soll durch Vergleich mit manuell erstellten Abkürzungsverzeichnissen (z.B. http://www.babylonia.org.uk) bewertet werden.

Einzelaufgaben

1. Aufbereitung des Datenmaterials. Als Ausgangsmaterial werden (noch unter Vorbehalt) pdf-Dateien verwendet, die in Text-Dateien überführt werden müssen (Verwendung von vorhandenen Werkzeugen).

2. Definition einer Heuristik für die syntaktischen Formen von Abkürzungen (z.B. Großbuchstaben, erstes Vorkommen in Klammern nach dem abgekürzten, evtl. aus mehreren Wörtern bestehenden Begriff)

3. Bestimmung von Position und Häufigkeit der Abkürzungen in den Text-Dateien

4. Entwicklung eine Algorithmus (z.B. Dynamische Programmierung), der Abkürzungen den abgekürzten Begriff zuordnet.

5. Aufbau des Lexikon mit Web-fähiger Benutzeroberfläche.

6. Bewertung der Effektivität des Verfahrens, insbesondere Bewertung der Präzision (precision).

Die Implementierung erfolgt wahlweise in Java, C# oder C. Zu empfehlen ist die Verwendung vorhandener Werkzeugen wie der Unix text tools.

Ein System der beschriebenen Art für die Bioinformatik ist der *abbreviation server* http://abbreviation.stanford.edu/ .

Maximale Teilnehmerzahl: 2

Voraussetzungen: Vordiplom, gute Kenntnisse von Algorithmen und Programmierung, Interesse an der Verarbeitung von Texten in natürlicher Sprache.

Betreuung: Schweppe

Das Projekt kann zu einer Diplomarbeit ausgebaut werden.