Klassifizierung von biochemischen Texten mittels statistischer Lernverfahren

Exposé einer Bachelorarbeit

Sebastian Schmeier

23. April 2003

Betreuer	Dr. Edda Klipp Max-Planck-Institut für Molekulare Genetik, Berlin Prof. Dr. Ulf Leser Institut für Informatik, Humboldt-Universität zu Berlin
Zeitraum	5. Mai - 30. Juni 2003

Motivation

Um biologische Systeme in silico zu modellieren, sind kinetische Daten unerlässlich. Diese Daten werden in aufwendigen experimentellen Messungen gewonnen und in natürlichsprachlichen Texten veröffentlicht. Die einzige Möglichkeit, solche Daten in Erfahrung zu bringen, besteht in zeitaufwendiger Literaturrecherche. Die existierende und täglich hinzukommende Masse an Literaturquellen ist allerdings längst nicht mehr überschaubar.
Dieser Umstand macht es sinnvoll, Abläufe zu identifizieren, die eine Klassifizierung von biochemischen Texten nach deren Relevanz bezüglich der Extraktion kinetischer Daten vereinfachen. Nach erfolgreicher Klassifizierung erhält man eine Aussage darüber, welche Dokumente interessante Informationen zu der gegebenen Problemstellung enthalten, kinetische Daten aufzufinden. Diese Arbeit gliedert sich in ein Projekt am Max-Planck-Institut für Molekulare Genetik ein mit dem Ziel, mit Hilfe der so gewonnen kinetischen Daten *Saccharomyces cerevisiae* in silico zu modellieren.

Zielsetzung

Ziel dieser Arbeit wird es sein, einen Prozess zu identifizieren und zu implementieren, welcher beliebig neue Texte bezüglich ihrer Relevanz für kinetische Modelle klassifiziert. Die Güte des Verfahrens wird anhand vorhandener, manuell annotierter Testdatensätze geprüft.

Vorgehen

Zum Erlernen einer Klassifikation und Unterscheidung zwischen relevanten und irrelevanten Dokumenten wird ein Trainingsdatensatz benötigt. Dieser ist vorher von Hand zu annotieren, d.h. jedes Dokument ist einer der beiden Klassen zuzuordnen.

Jedes Dokument lässt sich eindeutig durch einen sogenannte Wortvektor repräsentieren. Diese Darstellungsform erlaubt es, zwei Dokumente auf einfache Art und Weise miteinander vergleichbar zu machen. Die Gewinnung solcher Vektoren erfolgt mittels Verfahren zur Verarbeitung natürlicher Sprache *(Natural Language Processing, NLP)*. Jeder Text muss vorverarbeitet werden, um einzelne Worte und auch deren Grundformen zu bestimmen. Zum Einsatz kommen dabei Tokenization, Part-of-Speech-Tagging und Stemming. Weitere Verfahren sorgen für eine Gewichtung des Wortvektors unter Einbeziehung von relativen und absoluten Worthäufigkeiten in der gesamten Dokumentenbasis beziehungsweise in einzelnen Dokumenten. Als Methode zur Klassifizierung kommt ein Verfahren aus der statistischen Lerntheorie zum Einsatz, die sogenannten *Support Vector Machines*.
Einige der erwähnten Methoden und Implementationen von Algorithmen sind bereits verfügbar und sollen in verschiedenen Ausprägungen und Parametrisierungen zum Einsatz kommen. Eine Bewertung der unterschiedlichen Verfahren erfolgt anhand des manuell annotierten Testdatensatzes bezüglich Spezifität und Sensitivität der jeweils vorgenommenen Klassifizierung.