

Strings&Structures

Rolshoven, Jürgen

rols@spinfo.uni-koeln.de
Universität zu Köln, Deutschland

Etimi, Valmir

vetemi@smail.uni-koeln.de
Universität zu Köln, Deutschland

Seipel, Peter

pseipell@uni-koeln.de
Universität zu Köln, Deutschland

Wiehe, Thomas

twiehe@uni-koeln.de
Universität zu Köln, Deutschland

Der vorliegende Beitrag befasst sich mit dem Projekt "Strings&Structures. Codes of Sense and Function in Genomics and Linguistics". Dieses Projekt wird von der Sprachlichen Informationsverarbeitung und der Bioinformatik im Rahmen der Exzellenzinitiative der Universität zu Köln durchgeführt. Beide Bereiche befassen sich intensiv mit der Prozessierung von Texten. Bei der Sprachlichen Informationsverarbeitung handelt es sich um natürlichsprachliche Texte und Textkorpora, bei der Bioinformatik um genomische Texte. Das Projekt zielt auf die Aufdeckung von Mustern in Texten und die Analyse der Beziehung der Muster untereinander. Vor dem Hintergrund dieser Fragestellungen werden gemeinsam nutzbare Algorithmen entwickelt. Jedoch sollten dabei wesentliche Unterschiede der zugrundeliegenden Textarten nicht übersehen werden. Natürlichsprachliche Texte sind das Resultat grammatischer Produktionssysteme, genomische Texte sind Produktionssysteme. Das linguistische Vorhaben zielt auf die Rekonstruktion der erzeugenden Produktionssysteme aus zugrundeliegenden Textkorpora. Weitere Unterschiede zwischen natürlichsprachlichen Texten und genomischen Texten liegen in der Größe der zugrunde liegenden Alphabete und der zweigliedrigen Kombinatorikebenen natürlicher Sprachen. Wenngleich die Interaktion und Dynamik der Einheiten in genomischen Texten hochkomplex ist, so kann die Funktion einer einzelnen Einheit gut bestimmt werden. In natürlichen Sprachen dagegen ist die Bedeutung einzelner Einheiten oftmals nur schwierig zu bestimmen. Sie ist hochgradig kontext- und situationsabhängig. Dies hängt auch damit zusammen, dass sprachliche Einheiten weitgehend polysem sind.

Die automatische Aufdeckung der Bedeutung und Funktion sprachlicher Zeichen vollzieht sich in einem vierstufigen Prozess:

1. Ermittlung minimaler bedeutungs- oder funktionstragender Einheiten.
2. Kombinatorik dieser Einheiten durch Aufdeckung morphologischer Prozesse.
3. Syntaktische Kombinatorik der morphologisch erkannten Einheiten.
4. Auswertung syntaktischer Strukturen für die Bestimmung der Bedeutung sprachlicher Einheiten.

Dieses vierstufigen Verfahren wird in schrittweiser Verfeinerung in weitere Komponenten zerlegt, die algorithmisch als Module in einem Prozesskettensystem frei verschaltet werden. Ein solches graphisch orientiertes System ermöglicht auch Laien, Prozessketten für die Lösung eigener Fragestellungen zu schaffen.

Ad 1. Ermittlung minimaler bedeutungs- oder funktionstragender Einheiten.

Bei der Ermittlung minimaler Bedeutung oder funktionstragende Einheiten wird von dem strukturalistischen Grundgedanken der Zeichenkonstitution durch Opposition ausgegangen. Dieser Gedanke wird mit Hilfe von Suffixbäumen umgesetzt. In Suffixbäumen verweisen Verzweigungen auf potenziell in Opposition stehende Zeichenketten hin. Allerdings führt eine direkte Auswertung von Verzweigungen Suffixbäumen zu einer viel zu mächtigen Menge potentieller Morpheme. Daher müssen Filtermechanismen für deren Reduktion konstruiert werden. Ein Filtermechanismus beruht darin, nur identische Zeichenketten aus vorwärts und rückwärts aufgebauten Suffixbäumen zu verwenden.

Ad 2. Kombinatorik dieser Einheiten durch Aufdeckung morphologischer Prozesse.

Ein weiterer Filtermechanismus liegt in der Begrenzung der Kombinatorik von kleinsten funktions- oder bedeutungstragenden Einheiten. Formal kann morphologische Kombinatorik als Typ-2-Sprache im Sinne der Chomsky-Hierarchie formaler Sprachen betrachtet werden. Mit zusätzlichen Kriterien zur Unterscheidung bedeutungs- oder funktionstragender Einheiten kann die Übermenge, die der Suffixbaumgenerator liefert, drastisch reduziert werden. Die verbleibenden Einheiten sind in den nachfolgenden Schritten syntaktisch und semantisch zu analysieren.

Ad 3. Syntaktische Kombinatorik der morphologisch erkannten Einheiten

Eines der Probleme maschineller syntaktischer Sprachverarbeitung liegt in der Kontextsensitivität natürlicher Sprachen. Dies hat unter anderem zur Folge, dass Einheiten, die bedeutungsmäßig zusammengehören, in Sätzen oftmals weit voneinander getrennt sind. Für die Erkennung semantischer Zusammengehörigkeit und semantischer Abhängigkeit werden in dem vorliegenden Projekt Kookurrenzmatrizes ausgewertet. Die Kookurrenzmatrizes speichern semantische Vektoren, die semantische Abhängigkeit ausdrücken. Starke semantischer Abhängigkeit -etwa eines Verbs zu seinem Objekt -können werden in einer Baumstruktur direkt durch

benachbarte Knoten ausgedrückt, selbst dann, wenn es Vorkommen des Objekts gibt, die gar nicht unmittelbar neben dem Verb im Textkorpus stehen. Letztlich könnten auf diese Weise kontextabhängige Phänomene aufgedeckt werden.

Ad 4. 4. Auswertung syntaktischer Strukturen für die Bestimmung der Bedeutung sprachlicher Einheiten.

Syntaktische Strukturen in natürlichen Sprachen haben die Funktion, die Prozessierung sprachlichen Inputs zu erleichtern und zu beschleunigen. Syntaktische Strukturbäume ermöglichen es, korrekte Beziehungen zwischen sprachlichen Elementen herzustellen. Für die Bestimmung von Bedeutungspotenzial sind syntaktische Strukturen daher von grundlegender Bedeutung. Wird das Bedeutungspotenzial wiederum durch vektorielle Kookurrenzmatrizes erfasst, dann tragen syntaktischer Strukturbäume dazu bei, die Zahl der Komponenten der Matrizes stark zu reduzieren und folglich die vektorielle Semantik zu schärfen.

Eine Besonderheit des hier gewählten Vorgehens liegt in der Interaktion von subsymbolischen, vektoriellen und symbolischen, baumstrukturorientierten Verfahren. Die Stärke symbolischer Verfahren liegt in ihrer Kompaktheit und der Möglichkeit der Falsifikation. Subsymbolischer Verfahren sind nicht oder nur schwierig falsifizierbar. Sie machen semantische Unschärfe und semantische Ähnlichkeit fassbar