Zusammenfassung der Arbeit „Elements of Knowledge-free and Unsupervised Lexical Acquisition“ von Stefan Bordag

In dieser Arbeit wird ein Modell zur automatischen und unüberwachten Sprachwissensakquisition aus nicht-annotierten Sprachkorpora entworfen. Es werden weiterhin einige Algorithmen aus dem Modell heraus abgeleitet und an praktischen Problemstellungen getestet. Das Modell beruht dabei auf Ideen aus dem Strukturalismus in der Sprachwissenschaft. Es versucht einen vollständig berechenbaren Rahmen für die Entwicklung von Algorithmen zu geben, welcher auf einer möglichst kleinen Menge von universalen Prinzipien basiert. Von diesen Prinzipien wird angenommen, dass sie sich lediglich im Detail abweichend durch alle Ebenen der Sprache ziehen.
Die hypothetisierten Prinzipien sind Komposition von Einheiten zu komplexeren Einheiten und Abstraktion mit dem Effekt der bedingten Austauschbarkeit. Dementsprechend fußt das Modell auf den Mechanismen der Kookkurrenzbeobachtung, der Kontextvergleiche und Verallgemeinerungen. Es wird demonstriert, wie diese eigentlich einfachen Mittel zu komplexeren Konstrukten zusammengefügt werden können. Das wiederum ermöglicht das Formulieren von Algorithmen, welche spezifische Informationssorten extrahieren können. Dazu gehören unter anderem semantische Wortähnlichkeit, Mehrdeutigkeit, semantische Relationen, Morphologie, grammatische Kategorien und anderes.
In der zweiten Hälfte der Arbeit, ab Kapitel 3, werden existierende Verfahren zu einigen relevanten Problemstellungen überprüft und neue Lösung entsprechend dem Modell hergeleitet. Zunächst wird dabei auf Kookkurrenzbeobachtungen und semantische Wortähnlichkeitsberechnungen eingegangen. Unter Einführung einer neuen Testmethode wird gezeigt, welche Herangehensweisen besser geeignet sind und welche Kombination von Maßen die brauchbarsten Ergebnisse produziert. Die daraus gewonnenen Erkenntnisse bilden die praktische Grundlage für die weiteren Teile der Arbeit.
In dem darauf folgenden Kapitel wird eine eingangs getroffene Vereinfachung bezüglich der Kookkurrenzbeobachtungen revidiert. Ursprünglich wurde von jeder beobachteten Wortform angenommen, sie hätte nur eine Bedeutung. Werden jedoch die signifikanten Kookkurrenzen eines Wortes einem eigens dafür entwickelten Clusterverfahren unterzogen, zeigt sich, dass diese Vereinfachung oft fehlerhaft ist. Gleichzeitig kann mit Hilfe einer ebenfalls dafür entwickelten neuen Testmethode gezeigt werden, dass die Clusteranalyse intuitive und präzise Ergebnisse liefert.
Kapitel 5 der Arbeit befasst sich mit dem Phänomen, dass die meisten sprachlichen Ebenen, wie die Ebene der Phrasen oder der Morphologie, nicht direkt beobachtbar sind. Ausgehend von der Annahme, dass gleiche grammatische Merkmale oft mit gleichen sprachlichen Mitteln markiert werden, wird hier ein Verfahren entwickelt, welches die morphologische Ebene beobachtbar machen kann. Evaluiert wird dieses Verfahren im Rahmen eines internationalen Wettbewerbs derartiger Algorithmen. Darauf aufbauend wird ein Verfahren entwickelt, welches Morphemanalyse leistet, indem es verschiedene Morphe als das gleiche Morphem repräsentierend erkennt. Gleichzeitig wird damit gezeigt, dass in der Tat die gleichen Prinzipien der Komposition und Abstraktion auf unterschiedlichen sprachlichen Ebenen existieren.

Kapitel 6 schließt die Arbeit ab, indem es zeigt, wie auch zwischen semantischen, wie etwa symmetrischen und hierarchischen paradigmatischen Relationen, unterschieden werden kann. Unter Ausnutzung der in Kapitel 3 eingeführten Testmethoden werden einige Experimente durchgeführt, die die Effektivität der entwickelten Algorithmen belegen.
Insgesamt stellt diese Arbeit einen wichtigen Beitrag zur Forschung an sprachwissensfreien und unüberwachten Verfahren der Sprachwissensextraktion dar. Es wird anhand einer Ausformulierung eines neuen Modells, sowie an einer Vielzahl von praktischen Lösungen gezeigt, dass dieser gesamte Ansatz sehr viel versprechend ist und großes Potential für die weitere Forschung birgt.