Zusammenfassung: Markov Chain Monte Carlo Sampling for Dependency Trees

Christoph Teichmann

16. Mai 2014

In dieser Arbeit werden Markov Chain Monte Carlo (MCMC) Methoden für das Sampling von Dependenzbäumen entwickelt. Dependenzbäume sind ein Formalismus für die syntaktische Annotation von natürlichsprachlichen Sätzen. Dieser Formalismus hat in den letzten Jahren immer mehr an Bedeutung gewonnen und zur gleichen Zeit sind die Modelle, die zur Generierung und Verarbeitung von Dependenzbäumen verwendet werden, immer komplexer geworden.

Diese neuen Modelle machen es notwendig, Approximationsalgorithmen zu verwenden um Erwartungswerte und Optima zu berechnen. Beides wird häufig für Machine Learning Ansätze benötigt. Eine Klasse solcher Algorithmen sind die sogenannten Markov Chain Monte Carlo Techniken. Da Dependenzbäume eine starke innnere Struktur haben, besonders wenn linguistisch motivierte Teilmengen betrachtet werden, können sie nicht ohne weiteres durch MCMC Methoden bearbeitet werden. Die vorliegende Arbeit entwickelt Techniken, die die Anwendung von MCMC für Modelle von Dependenzbäumen unter sehr generellen Umständen ermöglichen. Es werden sowohl lokale Methoden entwickelt, die einen kleinen Teil eines Dependenzbaumes in einem Schritt verändern, als auch globale Ansätze, die eine komplette Neustrukturierung erlauben.

Nach der Vorstellung der verschiedenen Algorithmen werden diese anhand von zwei Problemen evaluiert. Das erste ist ein künstliches Problem, welches eine tiefere Einsicht in die verschiedenen Methoden erlaubt. Das zweite ist ein unüberwachtes Lernproblem, welches untersucht wie sich die Techniken bei größeren, untereinander verbundenen Daten verhalten.