Mehrwortausdruckserkennung mit phi-subsequentiellen longest-match-Transduktoren

Thomas Hanneforth
Linguistik Kolloquium Bielefeld, 2.6.2010

Bei der maschinellen Sprachverarbeitung wird die *Vorverarbeitungsphase* -- das Zergliedern des Eingabetextes in Einheiten, die dann an morphologische Analyse und Parser weitergegeben werden -- traditionell unterschätzt. Kern der Vorverarbeitung ist die effiziente Erkennung von Mehrwortausdrücken (MWA) wie "Rio de Janerio" und "zwischen Tür und Angel". Die Standardantwort darauf ist die Verwendung eines *Lexers* wie *flex* oder *re2c*, die auf der Grundlage einer Spezifikation effizienten C-Code erzeugen, mit dem dann die Eingabe deterministisch verarbeitet wird. Für große MWA-Listen stößt diese Vorgehensweise allerdings an ihre Grenzen, da leicht C-Code im Umfang von mehreren Millionen Zeilen generiert werden kann.

Das Problem dieses Ansatzes -- wie auch des verwandten Ansatzes von Mihov und Schulz (2005) -- ist, dass der unterliegende Automat immer eine *vollständige* Übergangsfunktion aufweist. Dies bedeutet, dass jeder Zustand für jedes Eingabesymbol einen wegführenden Übergang besitzt. Dies kann bei großen Eingabealphabeten zu einer Vervielfachung der Automatengröße führen.

Eine andere Antwort auf die Frage, wie man Wörter ohne vollständige Übergangsfunktion deterministisch erkennt, haben Aho & Corasick (1975) gegeben. Sie erweitern einen einfachen endlichen Automaten, der die Suchwörter repräsentiert, um eine sog. *failure-Funktion* $\phi: Q \rightarrow Q$, die jeweils einen Zustand des Automaten auf einen anderen Zustand abbildet. Die Funktion ϕ wird immer dann verwendet, wenn die Verarbeitung des aktuellen Symbols im aktuellen Zustand scheitert. Aho und Corasick haben gezeigt, dass dadurch die Verarbeitung eines Eingabetextes der Länge n im schlimmsten Fall $2n$ Schritte benötigt, also immer noch deterministisch ist.

Im Vortrag stelle ich einen Ansatz vor, der Aho/Corasicks Methode von der reinen Erkennung auf die *Transduktion*, also das Erzeugen einer Ausgabe, erweitert. Um hierbei Ambiguitäten zu vermeiden, verfährt der Transduktor nach der longest-match-Strategie, bei der beim Vorliegen von Suchwörtern, die an der gleichen Position beginnen, das mit der größeren Ausdehnung nach rechts präferiert wird (vgl. "New York" vs. "New Yorks").