Induktives Lernen von formalen Grammatiken durch Identification by Enumeration

FACHHOCHSCHULE ERFURT UNIVERSITY OF APPLIED SCIENCES
Angewandte Informatik

Betreuer: Prof. Dr. Oksana Arnold, Prof. Dr. Klaus P. Jantke

Studiengang Angewandte Informatik, Altonaer Str. 25, 99085 Erfurt, Tel. 0361 6700 642, e-mail: informatik@fh-erfurt.de

Lukas Bachmann

1995	Geboren in Bad Friedrichshall
2005-2011	Otto Klenert Realschule in Bad Friedrichshall
2011-2014	Gustav von Schmoller Schule in Heilbronn
2016-2022	Studium FH-Erfurt Master Angewandte Informatik

Identification by Enumeration

Identification by Enumeration ist ein induktiver Lernprozess. Dieser nimmt Informationen entgegen und leitet aus diesen eine allgemeingültige Beschreibung ab.
Dafür wird ein Generator benötigt, welcher Hypothesen der Reihe nach aufzählt. Die Hypothesen werden nacheinander betrachtet und darauf überprüft, ob sie gegenüber den bekannten Informationen konsistent sind. Inkonsistente Hypothesen werden verworfen, während das Aufzählen einer passenden Hypothese - und somit einer allgemeingültigen Beschreibung - zur Terminierung des Lernprozesses führt (vgl. Abb. 1).

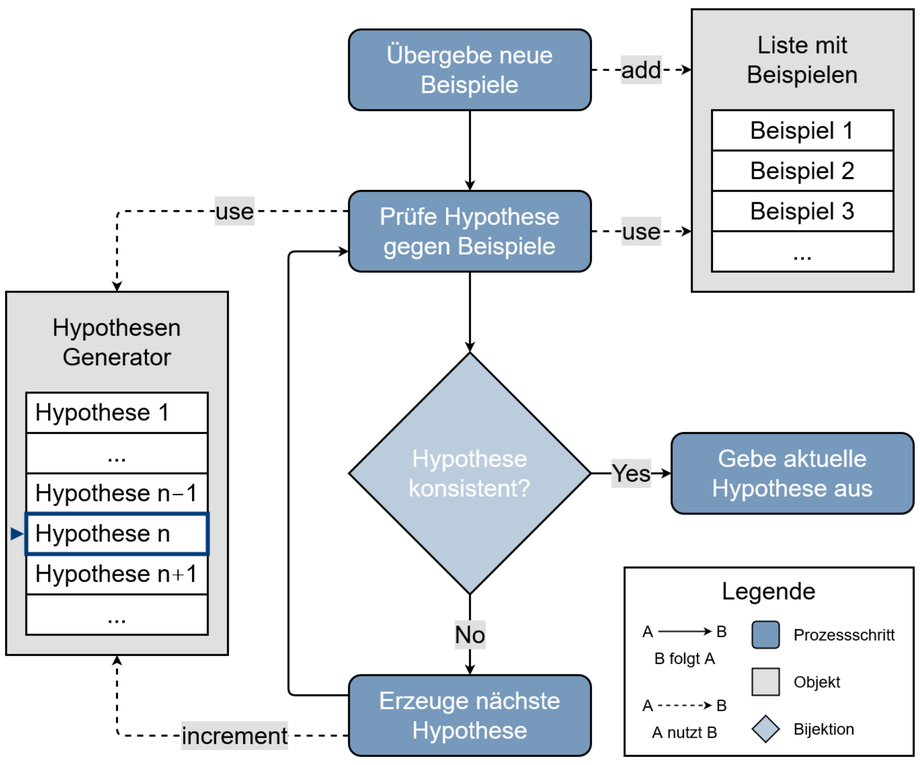

Abbildung 1: Darstellung des Lernprozesses

Grammatik *G* für die Sprache a^+b^+

$G = (N, \Sigma, P, s)$ mit

$N = \{s, n_1, n_2\}$

$\Sigma = \{a, b\}$

$P = \{s \rightarrow a\ n_1,\ n_1 \rightarrow a\ n_1,$

$n_1 \rightarrow b\ n_2,\ n_2 \rightarrow b\ n_2,$

$n_2 \rightarrow \varepsilon\}$

Akzeptor *A* für die Sprache a^+b^+

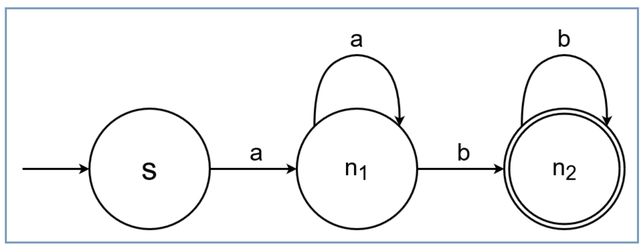

Abbildung 2: Überführung einer Grammatik in einen Akzeptor

Umsetzung

Ziel der Umsetzung ist es, rechtslineare Grammatiken und Grammatiken in Chomsky-Normalform zu lernen. Als Informationen sollen dem Lernprozess Wörter übergeben werden. Diese können Positiv- oder Negativbeispiele sein. Ein positives Beispielwort liegt in der Sprache, die die zu lernende Grammatik erzeugt. Negative Beispielwörter liegen nicht in dieser Sprache.
Als Hypothesen müssen formale Grammatiken durch den Generator aufgezählt werden. Dafür verfügt der Generator über eine Wissensbasis, in der das Alphabet, mit dessen Symbolen die Beispielwörter gebildet werden, und die Bildungsvorschrift für das Erzeugen der Grammatiken abgebildet ist.
Die Erzeugung der Grammatiken wird durch die Nutzung mehrerer Aufzählungen realisiert. Jede der Aufzählungen generiert eine Komponente einer Produktionsregel als Rückgabewert. Diese fließen kaskadenartig zusammen um vollständige Produktionsregeln zu erhalten. Aus den Produktionsregeln werden anschließend Mengen gebildet, welche die Hypothesen darstellen.
Um die Konsistenz einer Hypothese gegenüber den Beispielwörtern festzustellen, muss das Wortproblem gelöst werden. Kann mit einer Hypothese ein negatives Beispielwort gebildet bzw. ein positives Beispielwort nicht gebildet werden, ist sie nicht konsistent. Analog dazu kann mit einer konsistenten Hypothese jedes positive Beispielwort und keines der negativen Beispielwörter abgeleitet werden.
Um das Wortproblem für rechtslineare Grammatiken zu lösen, wird die Grammatik in einen Akzeptor überführt und anschließend das Beispielwort in diesen eingegeben (vgl. Abb. 2). Bei Grammatiken in Chomsky-Normalform wird die open-source Bibliothek Natural Language Tool Kit (NTLK) genutzt.

Ergebnisse

Mit dem implementierten Lernprozess konnten rechtslineare Grammatiken und Grammatiken in CNF für einfache a-b-Sprachen gelernt werden. Für Grammatiken von komplexeren Sprachen und Sprachen mit größerem Alphabet ist das Verfahren jedoch ungeeignet, da die benötigte Rechenzeit stark zunimmt.
Außerdem wurde festgestellt, dass durch die Eingabe ausschließlich positiver Beispielwörter immer eine Allsprache gelernt wird. Erst durch das Eingeben von Negativbeispielen kann diese weiter differenziert werden.