

Samenvatting

In de logica en de filosofie wordt onder de term inductie het proces verstaan waarbij natuurlijke wetten worden afgeleid door middel van het redeneren van onderdeel naar geheel, van het bijzondere geval naar het algemene of van het individuele naar het universele. Inductief redeneren voegt het totaal aan alle observaties of voorkomens samen tot een kortere wet die deze gebeurtenissen beschrijft en deductief kan voorspellen. Deductie is dan het redeneren vanuit tegengestelde richting: van de natuurlijke wet naar de observaties, van geheel naar onderdeel, van het algemene naar het bijzondere geval en van het universele naar het individuele.

Inductie tracht kennis te genereren, namelijk een hypothese die een verzameling empirische data of observaties verklaart binnen een kader van reeds aanwezige (achtergrond-)kennis. Inductief logisch programmeren is een machine learning principe dat inductief redeneren implementeert in het domein van logisch programmeren. Met andere woorden, gegeven een logisch programma B (achtergrondkennis) en D (data) proberen ILP algoritmes een logisch programma H (hypothese) te construeren zodat $B \wedge H \models D$.

In het typisch geval zal een ILP algoritme H opbouwen in stappen van één logische zin per iteratie met behulp van een incremental cover-strategie. Het zoeken naar de volgende logische zin wordt gedaan door een partiële ordening aan te brengen in de zoekruimte van alle mogelijke logische zinnen op een algemeen-specifiek-as. De operator die gebruikt wordt om deze ordening aan te brengen kan een generalisatie-operator zijn of een specialisatie-operator, wat respectievelijk resulteert in zoeken van specifiek naar algemeen of van algemeen naar specifiek. Het zoeken is aan de meest algemene kant gebonden door de lege, inconsistente zin \square , en aan de meest specifieke kant door de zogenaamde bottom clause, een minimale generalisatie van een positief voorbeeld. Progol is een dergelijk ILP algoritme en Aleph is een ILP systeem dat (onder meer) het Progol algoritme implementeert. Hoofdstuk 2 van dit proefschrift omvat een introductie in ILP in het algemeen en Progol en Aleph in het bijzonder.

Hoofdstuk 2 eindigt met een discussie over de keuze van ILP voor taalkun-

dige experimenten. ILP kent dezelfde voor- en nadelen als symbolisch rekenen in het algemeen: het vereist veel rekenkracht (zowel om een theory te construeren als om haar toe te passen) en het verwerken van numerieke data is lastig. Aan de andere kant, wanneer een symbolisch formalisme de voorkeur geniet, dan zijn logica en ILP de beste oplossing. Dat wil zeggen wanneer de uiteindelijke theorie niet alleen kwantitatief geëvalueerd en toegepast moet kunnen worden, maar ook de kwalitatieve analyse van de resultaten van belang is.

ILP is, zoals gezegd, een rekenkrachtintensieve taak, wat het een goed object voor parallelisatie maakt. Hoofdstuk 3 begint met een korte beschrijving van de *Message Passing Interface* (MPI). MPI is een specificatie van de *Application Programmers Interface* (API) voor message-passing libraries. Vervolgens wordt Aleph/MPI beschreven, een dataparallele versie van Aleph die ontwikkeld is voor de uitvoering van het onderzoek dat in dit proefschrift beschreven wordt. Aleph/MPI is gebaseerd op een uitbreiding van het YAP Prolog systeem met een interface (eveneens ontwikkeld voor dit project) naar MPI libraries. Hoofdstuk 3 sluit af met het testen en evalueren van deze uitbreidingen en enige speculatie over de domeinen waarin zij toegepast zouden kunnen worden.

De volgende twee hoofdstukken behandelen de toepassing van ILP op twee taalkundige domeinen: shallow parsing en fonotactiek. In hoofdstuk 4 worden de experimenten met shallow parsing besproken en in het bijzonder een experiment waarin getracht werd door middel van inductie een base NP chunker voor het Engels af te leiden. Een dergelijke chunker herkent alleen NP's van het laagste niveau (BaseNP). De NP 'confidence in the pound' wordt bijvoorbeeld geanalyseerd als:

(1) [NP [N1 Confidence] [PP in [N1 the pound]]]

maar de NP-chunker analyseert deze NP simpelweg als:

(2) [Confidence] in [the pound]

De invoer voor de chunker is platte tekst, die normaliter geannoteerd is met part-of-speech-informatie. Chunking is een probleem dat eindige toestandenautomaten aankunnen in één enkele run over de geannoteerde tekst.

Het experiment in hoofdstuk 4 is zodanig opgezet, dat de chunker een syntactische tag toewijst aan ieder woord. Deze tag geeft aan of het woord wel of niet onderdeel is van een BaseNP. Het predikaat waarvoor een hypothese afgeleid moet worden is de relatie tussen een woord in een bepaalde context en een syntactische tag. De linker context bestaat uit woorden met een part-of-speech-label én een syntactisch tag, de rechter context bestaat uit

woorden die alleen geannoteerd zijn voor part-of-speech: de richting waarin over de tekst gegaan wordt is van links naar rechts. De geannoteerde tekst is afkomstig uit de Penn Treebank, een geannoteerd en geparseerd Engels corpus.

Kwantitatief kan de geconstrueerde theory de resultaten van stochastische leermethodes niet evenaren. Kwalitatief gesproken heeft theory de leesbaarheid en modulariteit van symbolische theorieën. Dit kan worden toegeschreven aan verschillende oorzaken, zoals de ruis die inherent is aan de data, maar ook de ruis die veroorzaakt wordt doordat de opzet van het experiment een lange afstandsverschijnsel zoals syntaxis in de vorm van een lokaal verschijnsel forceert.

In hoofdstuk 5 wordt vervolgens de toepassing van ILP in het domein van de fonologie besproken, en meer in het bijzonder de fonotactiek: de regels die bepalen welke opeenvolgingen van fonemen in een bepaalde taal zijn toegestaan en welke niet. Een fonotactisch model is met andere woorden een model dat de non-woorden van een taal in twee groepen indeelt: systematische gaten, die nooit woorden van de taal hadden kunnen zijn, en toevallige gaten, die woorden hadden kunnen zijn, maar het toevalligerwijs niet zijn.

De algemene opzet is vergelijkbaar met die van het chunkingexperiment: het doel is een Nederlandse woordherkenner die eenmaal vanuit de nucleus naar buiten toe over de lettergreep heengaat en beslist of het een mogelijke lettergreep is. Het experiment zoals dat hier beschreven wordt is beperkt tot eenlettergrepige woorden, ervan uitgaande dat de fonotactiek in zijn geheel bestaat uit verschillende problemen: het herkennen van mogelijke lettergrepen en het combineren van lettergrepen tot welgevormde woorden.

Het doel is een predikaat dat een klinker of tweeklank en een gedeelte van het pre- of postvokale materiaal verbindt met een verzameling fonemen waarmee het kan combineren tot een welgevormde lettergreep. Het zij daarbij opgemerkt dat het niet nodig is aan te nemen dat alle tussenliggende stadia van een welgevormde lettergreep eveneens welgevormd zijn. De data zijn afkomstig uit het Nederlandse gedeelte van de CELEX Lexical Database. Twee verschillende manieren om de Nederlandse medeklinkers weer te geven zijn geprobeerd, de een uitgebreider en gedetailleerder dan de andere. De resultaten worden met elkaar vergeleken en hierbij blijkt dat de ILP leermethode profijt heeft van de uitgebreidere achtergrondkennis, waardoor een theorie genereerd wordt die compacter is en bovendien beter resultaten levert.

Het laatste hoofdstuk bevat een bespreking en samenvatting van de conclusies die aan het eind van elk hoofdstuk getrokken konden worden, maar daarnaast worden in dit hoofdstuk algemene conclusies getrokken uit het project als geheel. In het bijzonder wordt gesuggereerd dat chunking (of

tenminste het experimentele kader dat hier gekozen is) niet geschikt is voor de ILP leermethode, omdat chunking het moeilijk maakt om voordeel te halen uit de aantrekkelijke eigenschappen van symbolisch machine learning, namelijk de mogelijkheid om expliciete achtergrondkennis te gebruiken en de leesbaarheid van de resultaten. Wat betreft de fonotactische experimenten wordt gesuggereerd dat deze meer succesvol waren omdat de fonologische data minder ruis bevatte dan de treebank en omdat fonologie een lokaal probleem is, wat het gemakkelijker maakt om het in het experimentele kader van de hoofdstukken 4 en 5 te passen zonder informatie te verliezen.