Samenvatting

Dit proefschrift bespreekt de toepassing van leertechnieken bij de computationele verwerking van natuurlijke taal. We hebben drie verschillende leermethoden gebruikt voor het genereren van fonotactische modellen. Deze modellen hebben als taak het beoordelen van reeksen letters. Zij moeten kunnen beslissen of een bepaald woord wel of niet mogelijk is in een taal. Een goed fonotactisch model voor het Nederlands zal bijvoorbeeld 'kraag' goedkeuren en 'grmbl' afkeuren.

In ons onderzoek hebben we geprobeerd antwoorden te vinden op de volgende drie vragen:

1. Welke leermethode genereert de beste fonotactische modellen?
2. Heeft de representatiemethode van de leergegevens invloed op de kwaliteit van de geproduceerde fonotactische modellen?
3. Worden de fonotactische modellen beter als de leermethode taalkundige basiskennis beschikbaar heeft?

We hebben tien groepen leerexperimenten uitgevoerd. In elk van de experimentgroepen kregen de leermethoden ongeveer vijfduizend eenlettergrepige Nederlandse woorden aangeboden. Op basis van deze leerdata moesten zij een fonotactisch model voor eenlettergrepige Nederlandse woorden bouwen. Deze modellen zijn daarna getest met zeshonderd nieuwe eenlettergrepige Nederlandse woorden en zeshonderd letterreeksen die geen Nederlands woord vormen.

De eerste door ons geteste methode heet Hidden Markov Model (HMM). HMMs gebruiken een statistische leertechniek. De door HMMs geproduceerde fonotactische modellen presteerden goed. De beste HMMs accepteerden 99% van de correcte testdata en keurden 99% van de incorrecte testdata af. HMMs die startten met taalkundige basiskennis bouwden sneller een goed fonotactisch model dan HMMs die geen taalkundige basiskennis kregen aangeboden. De modellen van beide groepen HMMs presteerden ongeveer even goed.

Als tweede hebben we een neuraal netwerk met de naam Simple Recurrent Network (SRN) onderzocht. Dit netwerk presteerde erg slecht: het accepteerde alle correcte testdata maar keurde nooit meer dan 5% van de incorrecte testdata af. Dit resultaat heeft ons verbaasd omdat in vergelijkbare experimenten beschreven in de SRN-

literatuur alle correcte testdata wordt goedgekeurd en alle incorrecte testdata wordt afgekeurd. We konden aantonen dat het prestatieverschil ligt aan de complexiteit van de leerdata. Onze leergegevens uit het Nederlands zijn veel complexer dan de data die is gebruikt in de SRN-literatuur.

De derde leermethode die we hebben toegepast is een regelgebaseerde methode genaamd Inductive Logic Programming (ILP). Deze leertechniek produceerde ook goede fonotactische modellen. Het beste model gegenereerd door ILP accepteerde 99% van de correcte testdata en keurde 98% van de incorrecte testdata af. Er was een opvallende prestatieverbetering merkbaar in de ILP-experimenten waar taalkundige basiskennis beschikbaar was. Het afkeurpercentage voor incorrecte data lag voor modellen die waren gegenereerd zonder het gebruik van deze kennis tussen 60 en 70%. ILP met taalkundige basiskennis produceerde modellen die gemiddeld 98% van de incorrecte testdata afkeurden.

Na het uitvoeren van onze experimenten konden we onze onderzoeksvragen beantwoorden. Wat betreft de prestatie van de leermethoden: HMMs en ILP genereren fonotactische modellen die veel beter zijn dan de modellen geproduceerd door SRNs. Hoewel de ILP-modellen iets slechter presteren dan de HMM-modellen adviseren we het gebruik van ILP voor vervolgonderzoek. ILP heeft namelijk minder rekentijd nodig en de gegenereerde modellen bestaan, in tegenstelling tot HMM-modellen, uit regels die mensen kunnen interpreteren en manipuleren.

Voor het beantwoorden van onze vraag over datarepresentatie hebben we twee verschillende representatiemethoden vergeleken: de orthografische methode en de fonologische methode. In de eerste methode zijn woorden gerepresenteerd als reeksen van letters. In de tweede representatiemethode zijn woorden gecodeerd als reeksen van fonologische symbolen. Zowel HMM-modellen als ILP-modellen presteerden beter voor fonologische data dan voor orthografische data. SRNs zijn alleen maar toegepast op orthografische data.

De experimenten met taalkundige basiskennis leverden modellen op die op bijna alle punten beter presteerden dan de modellen die waren gebouwd zonder deze kennis te gebruiken. Het verschil was het grootst voor de ILP-modellen in de scores voor het afkeuren van incorrecte testdata. Het beschikbaar stellen van taalkundige basiskennis helpt leeralgorithmen dus bij het produceren van betere datamodellen.

We concluderen dat HMMs en ILP goede fonotactische modellen kunnen bouwen voor eenlettergrepige woorden uit een natuurlijke taal zonder dat incorrecte data hoeft te worden aangeboden tijdens het leerproces. SRNs genereren slechtere modellen omdat zij moeite hebben met de complexiteit van de data. De beste fonotactische modellen kunnen worden verkregen door de data te representeren als reeksen van fonologische symbolen en door tijdens de leerfase taalkundige basiskennis ter beschikking te stellen.