UNIVERSITEIT VAN AMSTERDAM

Taalverwerking met artificiële neurale netwerken
D. Hupkes

Artificiële neurale netwerken zijn opmerkenswaardig goed geworden als modellen voor verschillende natuurlijke taalverwerkingstaken. In dit proefschrift onderzoek ik of dit soort modellen daardoor ook gebruikt kunnen worden om meer over natuurlijke taal te leren. Ik focus specifiek op hiërarchische compositionaliteit in recurrente neurale netwerken (RNN).

Net zoals het menselijke taalverwerkingssysteem verwerken deze modellen binnenkomende signalen op een incrementele wijze, hun temporele structuur in acht nemend. Ik stel twee vragen:

a) Zijn RNN modellen in staat om hiërarchisch compositionele structuren correct te verwerken (gedragsmatige gelijkenis)?

b) Hoe kunnen we inzicht verkrijgen in de manier waarop ze dat doen (interpreteerbaarheid van modellen)?

Ik behandel deze vragen in zes hoofdstukken, die onderverdeeld zijn in drie delen. In deel één beschouw ik kunstmatige talen, wat mij in staat stelt om de verwerking van talige structuren in isolatie te bestuderen. In dit gedeelte introduceer ik ook diagnostische classificatie -- een interpreteerbaarheidstechniek die een belangrijke rol speelt in dit proefschrift -- en bezin wat het voor een model betekent om hiërarchische compositionaliteit te kunnen verwerken.

In deel twee bestudeer ik taalmodellen die getraind zijn op naturalistische data (Engelse zinnen). Eerder onderzoek wees uit dat dit soort modellen in staat zijn om syntaxgevoelige onderwerp-gezegde relaties te bevatten. Ik onderzoek hoe ze dat doen. Ik presenteer een gedetailleerde analyse van de interne dynamiek van modellen, waarvoor ik diagnostische classificatie, neuron ablatie en gegeneraliseerde contextuele decompositie gebruik.

Ten slotte onderzoek ik in het laatste gedeelte van dit proefschrift of de oplossing die een model vindt gestuurd kan worden door een aangepast leersignaal. Ik gebruik de technieken die eerder in het proefschrift geïntroduceerd zijn om de impact van dit aangepaste leersignaal in kaart te brengen.

Samenvattend presenteer ik in dit proefschrift verscheidene analyses die de geschiktheid van RNN modellen aangaande hierarchische syntactische structuur betreffen, evenals verschillende technieken die gebruikt kunnen worden om deze moeilijk interpreerbare modellen te begrijpen. De resultaten schetsen een positief beeld, dat suggeert dat RNN modellen wel degelijk nuttig kunnen zijn als verklarende modellen van natuurlijke taalverwerking.