Hoe de computerlinguïstiek gered werd door de taalkunde en omgekeerd

Walter Daelemans, Universiteit Antwerpen

Samenvatting

Hoewel de computertaalkunde is ontstaan als methode om taalkundige theorieën te testen, ging ze al snel haar eigen weg, gebaseerd op statistische methodes en zelflerende systemen. Tegenwoordig, met de "revolutie" van diepe neurale netwerken, met name de "pre-trained" taalmodellen, lijkt de taalkunde nog maar weinig relevant voor de computertaalkunde. In deze commentaar argumenteer ik dat dit een onterechte conclusie is: de kracht van deze neurale modellen is gebaseerd op een taalkundige theorie en omgekeerd kunnen de computermodellen een nieuwe impuls voor de taalkunde bieden.

Abstract

Although Computational Linguistics originated as a method for the evaluation of linguistic theories, it soon took a different path, guided by statistical and machine learning methods. With the current 'revolution' of deep neural networks, especially the pre-trained language models, linguistics doesn't seem relevant anymore for computational linguistics. In this comment, I argue that this conclusion is incorrect: the power of these neural models is based on a linguistic theory, and vice versa, current computational models can provide a new impulse to linguistics.

1. Taalkunde in de computerlinguïstiek

Bij de start ervan als aparte discipline in de jaren 60 was de computertaalkunde nog stevig ingebed in de taalkunde. De vader van de computertaalkunde in de Nederlanden, Hugo Brandt Corstius, hoewel opgeleid als wiskundige, ontwikkelde zijn computermodellen met als belangrijkste doel het toetsen van taalkundige theorieën: als je een taalkundige theorie implementeert als een computermodel en toepast op taaldata, maakt ze dan de juiste voorspellingen (Brandt Corstius, 1978)? Dat viel tegen, meestal was de taalkundige theorie niet specifiek, volledig of gedetailleerd genoeg om implementatie mogelijk te maken, laat staan de juiste voorspellingen te maken. Nochtans bleven de generatieve en functionele grammatica belangrijke inspiratie-

Verslagen en Mededelingen van de KANTL, vol. 130, nr. 1, 2020; 33-37

bronnen voor modelbouwers, zelfs als ze aan praktische toepassingen werkten zoals het Eurotra automatisch vertaalsysteem (European Commission, 1990) of systemen voor automatische zinsontleding. Dat bleef duren tot begin jaren 90, hoewel de taalkunde als inspiratiebron in de jaren 80 al concurrentie kreeg bij computertaalkundigen van een benadering die meer geïnspireerd was op de Kunstmatige Intelligentie van het moment, dus met meer nadruk op de representatie van niet-taalkundige domeinkennis en (logische) kennisrepresentatieformalismen, maar nog steeds compatibel met de meeste taalkundige theorieën. Er was één probleem: de aanpak was wel theoretisch elegant en interessant, maar werkte helaas niet. De computertaalkunde slaagde er niet in om robuuste en schaalbare toepassingen te ontwikkelen, en dat was tenslotte toch haar bestaansreden. Geïnspireerd door het relatieve succes van spraakherkenning en *information retrieval* geraakten computerlinguïsten hoe langer hoe meer gecharmeerd door de statistische methodes die daar werden gebruikt. Vanaf de start van de statistische revolutie in de computertaalkunde begin jaren 90 is de invloed van de taalkunde dan ook weggedeemsterd door het succes van statistiek en zelflerende systemen (*Machine Learning*, ML).

Het gebruik van zelflerende systemen in de computertaalkunde is eigenlijk een soort corpustaalkunde, maar dan zonder taalkundige: het systeem vindt autonoom patronen in de taaldata. Soms lijken die patronen op de regels die taalkundigen voorstellen, maar vaak is er weinig verband. Toch is het merkwaardig dat deze datageoriënteerde methodes grotendeels genegeerd werden door de taalkunde, behalve dan misschien in de computationele psycholinguïstiek (bijvoorbeeld Broeder & Murre, 2000, een bundel waar klassieke, handgemaakte, regelgebaseerde systemen en zelflerende methodes broederlijk naast elkaar staan). Nochtans beschrijven en verklaren deze statistische methodes veel beter dan regelgebaseerde alternatieven dat ‘grammatica een samenspel is van verschillende interne factoren’ en modelleren ze beter de onstabiele en manipuleerbare varianten van de ‘derdegolf-sociolinguïstiek’. De kwantitatieve wending in de taalkunde had veel vroeger kunnen starten. Het is trouwens niet onmogelijk om een exemplaar-gebaseerde grammatica te bedenken en implementeren die de fluïditeit van de taal kan verklaren (Skousen, 1989) en predicties kan maken die overeenkomen met de empirische data (Daelemans *et al.*, 1994). Dit is niet het uiteenvallen van de grammatica, maar een andere organisatie ervan. Exemplaar-gebaseerde theorieën worden wel genoemd door Van de Velde en van der Horst, maar verdienen een veel prominentere rol in de taalkunde dan ze toegewezen krijgen.

2. Diepe Neurale Netwerken

Zelfs tijdens de statistische revolutie bleef de taalkunde beperkt relevant. De zelflerende algoritmen van die tijd hadden informatie nodig als '*input features*' en die waren vaak taalkundig geïnspireerd. Taalkundigen bleven nuttig al was het maar voor de annotatie van data voor de zelflerende algoritmen (Hovy en Lavid, 2010).

Zelfs die rol wordt nu gespeeld door de algoritmen zelf. De revolutie van diepe neurale netwerken (vanaf ongeveer 2015), die niet alleen de computertaalkunde maar de hele Kunstmatige Intelligentie heeft overgenomen, wordt gekenmerkt door *end-to-end*-modellen (bv. een vertaalsysteem dat in één klap een zin in de brontaal naar de doeltaal omzet, zonder gebruik te maken van een syntactische of betekenis-representatie). Die modellen komen dan ook nog eens tot stand door zelforganisatie: de verschillende lagen van een diep neuraal netwerk specialiseren zich in specifieke subtaken zonder interventie van ontwikkelaars. Niemand hoeft na te denken over welke informatie eventueel nuttig zou kunnen zijn voor het zelflerende systeem. De meeste van die systemen hebben zelfs geen opsplitsing van de input in woorden meer nodig, maar ontwikkelen zelf hun morfologie op basis van patronen van letters of werken gewoon op letterniveau.

Tot grote ergernis van sommige taalkundig opgeleide computerlinguïsten kan een beetje informaticus met de alom beschikbare opensourcesoftware en data op een namiddag een vertaalsysteem maken dat beter werkt dan wat een team (computer)taalkundigen kon verwezenlijken op een taalkundig relevante manier in een meerjarig project.

De kracht van de aanpak is voor een groot deel gebaseerd op '*pre-trained language models*' (bv. BERT, Devlin *et al.*, 2018 en GPT, Brown *et al.*, 2020). Dit zijn taalmodellen die worden getraind op grote hoeveelheden tekst (onder meer) door het volgende woord te voorspellen gegeven voorgaande woorden of het middelste woord gegeven de context, en die via finetuning worden aangepast aan een specifiek domein of specifieke taak. Het resultaat van deze '*transfer learning*' is ongeziene accuraatheid en flexibiliteit. Deze taalmodellen zijn geen *one trick ponies* meer, maar kunnen ingezet worden voor heel uiteenlopende taalverwerkingstaken (vertaling, tekstproductie, vraag-antwoordsystemen, automatische samenvatting…). Je kan ook aantonen dat ze impliciet zowel syntactische als semantische kennis hebben. Bijvoorbeeld, in een taalmodel als BERT kan je rekenen met (de numerieke representaties van) betekenissen: koning – man + vrouw = koningin, Tokyo – Japan + België = Brussel enz. Je vindt er helaas ook de vooroordelen, impliciet in het

tekstcorpus waarop het taalmodel werd getraind, in terug: dokter – man + vrouw = verpleegster.

3. Wie redt wie?

Het is moeilijk om op dit moment te voorspellen of verdere ontwikkelingen binnen de huidige aanpak met diepe neurale netwerken zullen leiden tot oplossingen van resterende problemen voor de computertaalkunde. Die zijn er namelijk nog veel: impliciet en niet-letterlijk taalgebruik, beschikbaarheid van bredere wereldkennis en common sense, efficiëntere leerprocessen, en vooral: talige redeneerprocessen. Niemand weet of de nog steeds exponentieel groeiende rekenkracht en databeschikbaarheid zullen leiden tot oplossingen hiervoor, maar het is duidelijk dat de sprong voorwaarts, zowel in accuraatheid als in breedte (bv. multimodaliteit tussen beeld, spraak en tekst) kwalitatief is eerder dan incrementeel. Mijn stelling hier is dat we deze sprong te danken hebben aan de taalkunde. Taalmodellen als BERT en GPT zijn uiteindelijk niet meer dan operationaliseringen van de inzichten van Firth, Harris en anderen die aan de oorsprong liggen van de distributionele semantiek, die Van de Velde en van der Horst 'slechts een methode' noemen 'en geen inzicht'. Deze methode zou dan wel eens de belangrijkste bijdrage van de taalkunde aan de cognitiewetenschap ooit kunnen geweest zijn. In elk geval heeft ze de computertaalkunde gered, want geholpen door exponentieel groeiende rekenkracht en beschikbaarheid van data heeft deze methode voor de eerste keer gezorgd voor toepassingen met een grote accuraatheid en impact.

Of deze modellen ook cognitief relevant zijn is natuurlijk een andere vraag, maar ik vind het alleszins voorbarig om ze op dat vlak af te schrijven. Hoewel diepe neurale netwerken slechts een flauw wiskundig afkooksel zijn van biologische neurale netwerken, hebben ze op een voldoende abstract niveau toch dezelfde eigenschappen: globaal lokaal (specialisatie) en lokaal globaal (robuustheid), en met een belangrijke rol voor het geheugen bij taalverwerking (exemplaar-gebaseerd). Misschien kunnen deze modellen aanleiding geven tot een nieuwe taalkunde op basis van de analyse van getrainde diepe neurale netwerken (Hu *et al.*, 2020) en kan de computertaalkunde zo op zijn beurt de taalkunde redden.

Literatuurlijst

Brandt Corstius, H. (1978). *Computer-taalkunde*. Coutinho.

Broeder, P. en Murre, J. (2000). *Models of language acquisition*. Oxford University Press.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. arXiv preprint arXiv:2005.14165

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding arxiv:1810.04805

Daelemans, W., Gillis, S., & Durieux, G. (1994).' The acquisition of stress: A data-oriented approach'. *Computational Linguistics*, 20(3), 421-453.

European Commission (1990). *The European Community's Research and Development Project on Machine Translation*.

Hovy, E. and Lavid, J. (2010). 'Towards a "Science" of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics'. *International Journal of Translation*, Vol. 22, no 1.

Hu, J., Gauthier, J., Qian, P., Wilcox, E., & Levy, R. P. (2020). A systematic assessment of syntactic generalization in neural language models. arXiv preprint arXiv:2005.03692.

Skousen, R. (1989). *Analogical modeling of language*. Springer.