The automatic acquisition of a Dutch lexicon for opinion mining

Maks, E.

2018

document version
Publisher's PDF, also known as Version of record

Samenvatting

Dit proefschrift beschrijft het ontwerp en de automatische acquisitie van een Nederlands lexicon voor *opinion mining*. Met *opinion mining* bedoelen we het aan de oppervlakte brengen van opinies en ideeën die mensen hebben over allerlei onderwerpen: Is Rome de mooiste stad van de wereld? Is Brexit slecht voor de Europese economie? Moet je je kinderen wel laten vaccineren? Er zijn allerlei redenen waarom het interessant is te weten wat mensen hiervan vinden. Zo willen toeristenorganisaties weten waar mensen graag op vakantie gaan. Overheidsinstanties willen weten hoe mensen over vaccinaties denken. Politici willen weten hoe mensen over de Brexit denken. Er is een enorme hoeveelheid teksten in digitale vorm en meestal online beschikbaar waardoor we te weten kunnen komen welke opvattingen mensen hebben over deze onderwerpen.

Opinion mining is een vorm van automatisch tekstanalyse die zich richt op het zoeken van meningen in teksten. Over het algemeen wordt daarbij een computationeel lexicon gebruikt, bijvoorbeeld om vast te stellen of een woord een positieve of een negatieve betekenis heeft. Dit proefschrift presenteert onderzoek naar hoe een dergelijk lexicon eruit moet zien, hoe het kan worden gemaakt en hoe het kan worden ingezet bij *opinion mining* in het Nederlands. We hebben vier onderzoeksvragen onderscheiden die achtereenvolgens behandeld worden in de hoofdstukken 2 t/m 5 van het proefschrift.

Hoofdstuk 2 *Wat zijn de specificaties waaraan een computationeel lexicon voor opinion mining moet voldoen en hoe kunnen die het best worden geïmplementeerd?*
Om een antwoord op deze vraag te geven hebben we een overzicht gemaakt van de belangrijkste lexica en corpora die in het veld van *opinion mining* gebruikt worden. Het uitgangspunt is een theoretisch model (Martin and White (2005)) voor *attitudinal language* (d.w.z. de taal die mensen gebruiken om hun meningen en gevoelens te uiten). Met dit model hebben we bestaande corpora en lexica geanalyseerd. De analyse wees uit dat in een lexicon voor *opinion mining* in ieder geval de volgende aspecten van woordbetekenis moeten worden opgenomen: polariteit, *multiple actor attitude* en semantische categorie.

- Polariteit geeft aan of een woord een positieve of een negatieve betekenis heeft. In de volgende voorbeelden heeft *geweldig* een positieve betekenis, en *miste* een negatieve.

  (98) Dit was een geweldige vakantie

  (99) Hij miste de tijd dat hij nog kon doen en laten wat hij wilde.

Een *opinion mining*-systeem kan deze informatie gebruiken om een tekst te analyseren en met behulp van de frequentie van positieve en negatieve woorden vast te stellen of er een positief of negatief oordeel wordt uitgedrukt.

- Voor *opinion mining* is het niet alleen van belang om te weten of een mening positief of negatief is, maar ook **wie** die mening vertolkt. Ook dat kan afgeleid worden uit de betekenisstructuur van een woord. *Multiple actor attitude* verwijst naar het verschijnsel dat een woord de attitude van verschillende actoren kan uitdrukken: de attitude van de spreker of schrijver van een tekst (afgekort als SW - (speaker/writer)), van een actor die opgevoerd wordt in de tekst (afgekort als AC) of van beide. In voorbeeld (100) verwoordt een SW zijn negatieve houding ten aanzien van de 'hij' door het werkwoord 'zeuren' te gebruiken, in plaats van bijvoorbeeld het neutralere 'praten'. In voorbeeld (101) wordt een actor opgevoerd die een negatief oordeel heeft over een voorstel wat duidelijk wordt uit het gebruik van het woord 'verwerpen'. De SW is in dit geval neutraal. Voorbeeld (102) laat zien dat een woord als 'opscheppen' ook een meervoudige attitude kan uitdrukken: 'Hij', de actor, is positief over wat hij bereikt heeft, maar de SW is negatief over 'hij'.

  (100) Hij zeurt over van alles en nog wat (SW)

  (101) Hij verwerpt het voorstel (AC)

  (102) Hij schept op over alles wat hij bereikt heeft (SW/AC)

  In de rest van deze samenvatting wordt de term *multiple actor attitude* kort weergegeven met 'attitude'.

- Een indeling van woorden in semantische categorieën kan helpen bij *opinion mining* omdat sommige klassen van woorden een grotere rol spelen bij het verwoorden van meningen dan andere. Zo zullen woorden uit de categorie 'Behaviour' zoals bijvoorbeeld 'eerlijk' eerder een opinie uitdrukken dan een woord als 'donkerblauw' uit de categorie 'Color'.

De rest van het onderzoek had betrekking op de eerste twee van deze specificaties: polariteit en attitude.

Hoofdstuk 3 *Hoe kunnen de specificaties die in hoofdstuk 2 geformuleerd zijn, worden opgenomen in een lexicon en op een betrouwbare manier worden geannoteerd?*
In dit hoofdstuk wordt beschreven hoe we de bovengenoemde specificaties polariteit en attitude hebben geïntegreerd in een bestaand lexicon voor het Nederlands (Cornetto, Vossen et al. (2008)). We hebben een annotatieschema ontwikkeld en een set woorden geselecteerd die representatief zijn voor opiniërend taalgebruik. Twee onderzoekers hebben deze set onafhankelijk van elkaar geannoteerd. Een vergelijking wees uit dat de meeste aspecten van het annotatieschema betrouwbaar geannoteerd konden worden met een relatief hoge intercodeurbetrouwbaarheid. Door deze handmatige annotatie beschikten we bovendien over een gouden standaard die in het vervolg van het onderzoek gebruikt kon worden voor de evaluatie van de methodes voor automatische acquisitie van polariteit en attitude (zie hoofdstuk 4).

Hoofdstuk 4 *Wat zijn geschikte methodes om automatisch lexicons te genereren die voldoen aan bovengenoemde specificaties (hoofdstuk 2)?*
In hoofdstuk 4 worden 5 methodes beschreven en getest die polariteit en attitude identificeren in een lexicon of corpus. Twee hiervan richten zich op polariteit (positief of negatief) en de overige 3 op attitude (SW of AC):

- *cross-lingual transfer method* Deze methode identificeert polariteit door via de vertaallinks die bestaan tussen het Engelse lexicon SentiWordnet en het Nederlandse Cornetto polariteitswaarden over te brengen van het Engels naar het Nederlands.
- *wordnet propagation method* Deze methode identificeert polariteit in het wordnet-gedeelte van Cornetto. De methode gaat ervan uit dat synoniemen, hyponiemen, en antoniemen overeenkomen in polariteit. De methode start met een *seedlist* (een lijst met woorden waarvan de polariteit bekend is), verzamelt dan meer woorden in het wordnet en kent daar polariteit aan toe door gebruik te maken van links tussen synoniemen, hyperoniemen en antoniemen.
- *lexical feature method* Dit is een *machine learning*-methode die ervan uitgaat dat morpho-syntactische, syntactische, semantische en pragmatische eigenschappen van woorden zoals die beschreven zijn in een lexicon kunnen helpen bij het onderscheiden van de verschillende vormen van attitude.
- *corpus comparison method* Dit is een corpus-gebaseerde methode die ervan uitgaat dat de verdeling van woorden met verschillende soorten attitude verschilt per genre. We hebben een corpus samengesteld met drie verschillende genres: nieuwsartikelen, commentaar op nieuwsartikelen en Wikipedia-artikelen. De methode classificeert de woorden die relatief veel voorkomen in nieuwsartikelen als woorden met AC-attitude en woorden die relatief veel voorkomen in commentaar op nieuwsartikelen als woorden met SW-attitude. Woorden die relatief veel voorkomen in Wikipedia worden als neutraal beschouwd.
- *lexical pattern method* Deze methode gaat ervan uit dat woorden met dezelfde attitude vaak voorkomen in een vergelijkbare context. De methode start met lexicale patronen die een associatie hebben met woorden met ofwel AC-attitude ofwel SW-attitude. Met deze patronen worden nieuwe woorden in het corpus gezocht en ingedeeld in AC-attitude of SW-attitude.

De belangrijkste conclusies van hoofdstuk 4 zijn dat attitude moeilijker is te identificeren dan polariteit en dat AC-attitude moeilijker te identificeren is dan SW-attitude. Bovendien blijkt dat zowel attitude als polariteit beter identificeerbaar zijn bij bijvoeglijke naamwoorden dan bij zelfstandige naamwoorden en werkwoorden. Een andere conclusie is dat methodes die gebruik maken van linguïstische kennis zoals de *lexical feature* en de *lexical pattern* methode beter scoren dan de andere methodes.

Hoofdstuk 5 *Hoe kunnen de automatisch gegenereerde lexicons (zoals beschreven in hoofdstuk 4) gebruikt worden om de kwaliteit van opinion mining in het Nederlands te verhogen?*

Om een antwoord te geven op deze vraag hebben we een tweetal toepassingen geformuleerd:

- De eerste toepassing betreft de classificatie van hotel reviews als positief of negatief (*is het een geweldig of een afschuwelijk hotel?*). We voerden experimenten uit waarbij we verschillende classificatiemethoden toepasten en lieten zien dat het polariteitslexicon dat gegenereerd is met de *wordnet propagation*-methode de resultaten van de classificatie aanzienlijk verbetert.

- De tweede toepassing betreft de evaluatie van het attitudelexicon dat is gegeneerd door de *lexical feature*-methode. We implementeerden een *opinion mining*-systeem dat als doel heeft *opinion expressions* (wat is de opinie?), *opinion holders* (wie heeft die opinie) en *opinion targets* (waarover gaat de opinie) te identificeren in een geannoteerd corpus van nieuwsartikelen. We pasten een *machine learning*-benadering toe waarbij we SW- en AC-labels uit het lexicon als *features* gebruikten. Het experiment wees uit dat deze toepassing iets beter scoort met lexicon dan zonder lexicon, maar wel minder dan we hadden verwacht. Uit een nadere analyse van de redenen waarom de resultaten minder goed waren dan verwacht, bleek dat de verdeling van AC-woorden en SW-woorden over de verschillende woordklassen (zelfstandige naamwoorden, bijvoeglijke naamwoorden en werkwoorden) niet gelijk is. AC-attitude wordt hoofdzakelijk uitgedrukt door werkwoorden en SW-attitude hoofdzakelijk door bijvoeglijke naamwoorden. Dit leidt ertoe dat met een *machine learning*-benadering zoals wij hebben toegepast de SW- en AC-labels grotendeels samenvallen met de woordklasse-features en daardoor weinig toegevoegde waarde hebben.

In vervolgonderzoek zou onderzocht kunnen worden hoe het attitude-lexicon beter kan worden ingezet. Het zou kunnen dat de toepassing die we gekozen hebben, nl. een analyse van nieuwsartikelen, niet de meest geschikte is om de bijdrage van het attitude lexicon te testen. De SW- en AC-labels die in een dergelijk lexicon gespecificeerd worden, kunnen misschien beter benut worden in een sterker opiniërend genre zoals bijvoorbeeld redactionele commentaren.

Vervolgonderzoek zou ook kunnen kijken naar hoe de automatische acquisitie van met name het attitude-lexicon verbeterd kan worden. Het zou bijvoorbeeld zinvol kunnen zijn afzonderlijke methodes te ontwikkelen voor de identificatie van ofwel AC ofwel SW-attitude. In ons geval hebben we steeds dezelfde heuristieken toegepast om tegelijkertijd AC-attitude én SW-attitude te identificeren. Een uitzondering hierop is de *lexical pattern*-methode waarmee het wel mogelijk is specifieke patronen te formuleren voor de identificatie van de verschillende soorten attitude op zich, en die beter dan de andere methodes scoort bij de identificatie van AC-attitude. Een ander mogelijkheid om de kwaliteit van het attitudelexicon te verbeteren zou in het gebruik van semantische en ontologische kennis kunnen liggen. In de *lexical feature*-methode, die de beste methode bleek te zijn voor de identificatie van attitude, is dit soort kennis al opgenomen. Deze methode kan verder worden uitgebreid door gebruik te maken van een semantische classificatie.