Making sense of legal texts

de Maat, E.

Samenvatting

De regels van een moderne samenleving zijn vastgelegd in wetten, maatregelen en jurisprudentie. Deze regels vormen een specificatie van hoe de samenleving dient te zijn. Ze vertellen ons hoe de burgers in die samenleving zich horen te gedragen, en ook hoe de regering en ambtenaren van die samenleving zich horen te gedragen. Qua tekst verschillen deze specificaties van technische specificaties, aangezien ze zijn geschreven in natuurlijke taal in plaats van een formele taal of diagrammen. Echter, wetgeving is gebaseerd op *best practices* en ontwerppatronen. Daardoor is wetgeving veel regelmatiger dan andere teksten in natuurlijke taal, en verschilt het niet zoveel van technische specificaties.

Binnen de rechtsinformatica, de tak van de informatica die zich bezighoudt met juridische onderwerpen, wordt deze regelmatigheid van de wetgeving gebruikt als basis voor toepassingen die de toegankelijkheid van de wetgeving vergroten. Een veel voorkomende toepassing is een *portal* dat het voor gebruikers mogelijk maakt om wetgeving te doorzoeken door middel van sleutelwoorden, titels, etc. In dergelijke portals is de wetgeving niet opgeslagen als platte tekst, maar in een meer gestructureerd opslagformaat. Deze opslagformaten maken het mogelijk om te verwijzen naar een specifiek deel van een document. Ook bieden ze de mogelijkheid om metadata op te slaan. Deze metadata kan ook worden gebruikt om de gebruiker te helpen de informatie te vinden die hij nodig heeft.

Voor gebruikers zonder een juridische achtergrond zijn deze toepassingen echter onvoldoende behulpzaam. Deze gebruikers zijn vaak op zoek naar het antwoord op een specifiek juridisch probleem, in plaats van alleen juridische informatie. Voor deze gebruikers is een verwijzing naar de relevante regelingen dan ook onvoldoende. Daarom wordt een tweede soort toepassingen ontwikkeld: systemen die de gebruiker vragen naar zijn situatie, die regels toepassen en een antwoord geven. Deze systemen zijn niet alleen nuttig voor burgers die een antwoord zoeken op hun juridische vraag, maar kunnen ook ambtenaren helpen bij het verwerken van routine zaken (bijvoorbeeld bij het verstrekken van vergunningen).

Voor toepassingen van deze tweede soort zijn complete, formele en uitvoerbare modellen van de wet nodig. Het maken van deze modellen kost veel inspanning. Dit is een probleem dat zich bij veel kennisgebaseerde systemen voordoet, en wordt ook wel de *knowledge acquisition bottleneck* genoemd. Het onderzoek dat hier wordt gepresenteerd is er op gericht om het effect van deze bottleneck te verminderen door Nederlandse wetten automatisch te vertalen naar computermodellen.

Deze automatische vertaling wordt niet in één stap gedaan. Het proces is opgesplitst in een aantal stappen:

1. structuur herkennen: de structuur van de wet wordt herkend en gemarkeerd;
2. verwijzingen zoeken: alle verwijzingen in de wet naar andere juridische bronnen worden opgezocht en gemarkeerd;
3. zinnen classificeren: de zinnen in de wet worden geclassificeerd;
4. zinnen modelleren: van elke zin wordt een model gemaakt;
5. modellen integreren: de losse modellen per zin worden geïntegreerd tot één compleet model van de gehele wet.

De Aanwijzingen voor de Regelgeving vormen de basis voor het herkennen van structuurelementen in Nederlandse regelgeving (beschreven in hoofdstuk 2). In de aanwijzingen wordt een wet verdeeld in zes onderdelen: opschrift, aanhef, lichaam, slotformulier, ondertekening en bijlagen. Het lichaam is vervolgens verdeeld in artikelen, die op hun beurt kunnen bestaan uit leden, en die lijsten met lijstonderdelen kunnen bevatten. De artikelen kunnen worden gegroepeerd in paragrafen, afdelingen, titels, hoofdstukken en boeken.

Deze verschillende onderdelen willen we kunnen herkennen en identificeren. Dat wil zeggen: we willen weten welke delen van een document bij elkaar horen, en we willen ook weten dat ze samen artikel 12 of hoofdstuk II vormen.

Voor de aanhef, het slotformulier en ondertekening van een regeling worden sjablonen gebruikt. Deze sjablonen worden voorgeschreven in de richtlijnen. Dit betekent dat ze eenvoudig opgespoord en geïdentificeerd kunnen worden door te zoeken naar de vaste onderdelen van deze sjablonen. De meeste andere onderdelen van een regeling zijn gemarkeerd met een kop of een index. De kop bestaat uit de aanduiding van het niveau (zoals *artikel* of *hoofdstuk*), een index (zoals *4*, *II* of *5:128*) en eventueel een titel. Door deze koppen op te sporen kan het begin van een nieuw structuurelement (en het einde van het vorige element) worden gevonden. Daarna moeten de artikelen en leden nog worden opgedeeld in zinnen. Dit kan worden gedaan met reeds bestaande toepassingen.

Met behulp van de General Architecture for Text Engineering (GATE) is een prototype parser gebouwd, gebaseerd op bovenstaande ideeën. Deze parser is getest op tien verschillende wetten. Op artikelniveau en hoger werd de structuur van de wetten foutloos bepaald. Binnen 2% van de artikelen was er een fout binnen de zinnen, lijsten of leden. Bij de overige 98% werd ook de structuur van de artikelen zonder fouten bepaald. De grootste uitdaging bij het bepalen van de structuur van regelgeving is het herkennen van wijzigingsteksten. Deze wijzingsteksten kunnen structuurelementen bevatten die geen onderdeel zijn van de (wijzigende) wet. In Nederlandse regelgeving zijn deze teksten niet gemarkeerd met aanhalingstekens. Dit betekent dat de structuurelementen die deel uitmaken van een wijzigingstekst herkend moeten worden door de indices te vergelijken met die van de naastliggende elementen (aangezien de indices van de elementen in de wijzigingstekst meestal geen nette reeks zullen vormen de indices van de omliggende tekst).

Net zoals koppen volgen verwijzingen specifieke patronen, die gebruikt kunnen worden om ze op te sporen (zie hoofdstuk 3). Verwijzingen naar regelgeving worden gemaakt door middel van de (citeer)titel of de datum van ondertekening. Onderdelen van een document worden aangeduid door middel van het niveau (*hoofdstuk*, *artikel*, *onderdeel*) en de index. Een complete verwijzing bevat een aanduiding van het document waarnaar verwezen wordt, maar incomplete verwijzingen, zonder een dergelijke aanduiding, komen veel voor. Het doeldocument van dergelijke verwijzingen hangt af van de context waarin ze voorkomen. Binnen een wet verwijst een incomplete verwijzing vaak naar een ander deel van die wet, tenzij de verwijzing deel uitmaakt van een groep wijzigingsinstructies. In dat geval is het vaak een verwijzing naar een onderdeel van de wet die wordt gewijzigd. Verwijzingen kunnen ook naar

meerdere locaties verwijzen, door meerdere indices op te nemen (zoals *artikelen 12, 13 en 14*) of door ze te combineren in een reeks (*artikelen 12-14*).

De verwijzingen die bestaan uit een niveau-aanduiding en een index (of meerdere indices, of een reeks) kunnen eenvoudig worden opgespoord met patronen. Hetzelfde geldt voor verwijzingen die de datum van ondertekening gebruiken om te verwijzen naar een document. Daarentegen variëren citeertitels enorm. Deze kunnen daarom niet worden beschreven met een patroon. Om ze op te sporen is daarom een lijst met alle bestaande titels nodig.

Er is een prototype parser ontwikkeld gebaseerd op een dergelijke lijst van titels en patronen voor verwijzingen op basis van ondertekendatum en niveau-aanduiding en index. Deze parser is getest op zes Nederlandse wetten met meer dan 1.000 verwijzingen. De parser spoorde 97% van deze verwijzingen succesvol op, met nauwelijks fout-positieven.

In hoofdstuk 4 wordt de classificatie van zinnen in Nederlandse wetten behandeld. Er worden vijftien verschillende soorten zinnen onderscheden. Een deel hiervan zijn de eigenlijke normzinnen: rechten/permissies en plichten. Deze zinnen worden ondersteund door definities, die de termen definiëren die door de normzinnen worden gebruikt. Gerelateerd aan de definities zijn de fictiebepaling, die bepaalde concepten gelijk verklaren voor de toepassing van bepaalde normen. Van toepassing verklaringen geven aan dat bepaalde normen wel (of niet) van toepassing zijn in bepaalde omstandigheden. Strafbepalingen stellen vast wat de straf is voor het overtreden van een norm. Een publicatiebepaling is een specifieke verplichting die opdracht geeft bepaalde informatie te publiceren. Een waardebepaling berekent een waarde, die binnen een norm wordt gebruikt.

Naast deze regels, die de eigenlijke wet vormen, zijn er ook een aantal zinnen die over de wet zelf gaan. Dit zijn zinnen die de datum van inwerkingtreding of de citeertitel vastleggen, en zinnen die het recht (of de plicht) tot het stellen van aanvullende regels delegeren. Andere zinnen wijzigen bestaande regelgeving door nieuwe tekst in te voeren, tekst te vervangen of te verwijderen, of door tekstelementen te hernummeren. Door middel van scopebepalingen wordt de scope voor zulke wijzigingen gezet, waarmee de verwijzingen in de tekst versimpeld worden.

Elke soort zinnen maakt gebruik van specifieke signaalwoorden en taalpatronen. Deze signaalwoorden en patronen kunnen worden gebruikt om zinnen te classificeren. Er is slechts één uitzondering. Hoewel veel verplichtingen signaalwoorden gebruiken, zoals *moeten* en *verplicht*, is het ook gebruikelijk om een verplichting te beschrijven als een feit. Dit wil zeggen dat de verplichting wordt beschreven alsof de betreffende actie altijd gebeurt (in plaats van dat hij moet gebeuren). In een dergelijke zin mist een signaalwoord zoals *moeten*. Aangezien alleen (bepaalde) verplichtingen geen gebruik maken van een patroon, kunnen de patronen nog steeds gebruikt worden voor de classificatie van zinnen. Als een zin voldoet aan geen enkel patroon, dan kunnen we aannemen dat het een verplichting is.

Er is een classifier gebouwd die gebruik maakt van dergelijke patronen. Deze classifier is getest op achttien verschillende wetten met daarin 591 zinnen (lijsten buiten beschouwing latend). Van deze 591 zinnen waren er 157 verplichtingen die waren beschreven als een feit. De classifier was in staat om 91% van deze zinnen correct the classificeren.

Met betrekking tot het classificeren van tekst blijkt automatisch leren vaak succesvoller te zijn dan een kennisgebaseerde aanpak. Daarom is er, naast de classifier gebaseerd op patronen, ook een test uitgevoerd gebaseerd op automatisch leren, met Support Vector Machines (SVMs). Hierbij is gebruik gemaakt van een leave-one-out (LOO) aanpak met dezelfde dataset die bij het testen van de op patronen gebaseerde classifier is gebruikt. De LOO aanpak houdt in dat er een classifier wordt getraind op alle zinnen op één na. Deze laatste zin wordt vervolgens geclassificeerd met de gemaakte classifier. Dit wordt herhaald voor alle zinnen. Op basis van het aantal correcte classificaties kan vervolgens een LOO nauwkeurigheid worden bepaald. Dit is een voorspeller voor de nauwkeurigheid van de uiteindelijke classifier. Voor de classificatie van de zinnen in wetgeving haalde de aanpak gebaseerd op automatisch leren een nauwkeurigheid van 95%. In een test op twee nieuwe wetten classificeerde de automatisch leren classifier 95,77% respectievelijk 88,78% van de zinnen correct, vergeleken met 94,37% en 95,61% door de patroongebaseerde classifier. Dit suggereert dat geen van beide methodes strikt beter is dan de ander. De patroongebaseerde methode heeft wel tot voordeel dat het duidelijk is op basis van welke kenmerken de classificatie van een zin heeft plaatsgevonden.

De volgende stap is het eigenlijke modelleren, besproken in hoofdstuk 5. Bij het modelleren geldt dat er een belangrijk verschil is tussen de zinnen die gaan over de wet zelf en de zinnen die gaan over het onderwerp van de wet. De zinnen die over de wet gaan zijn de zinnen die bestaande wetten wijzigen, de datum van inwerkingtreding of de citeertitel van een wet vastleggen, etc. Deze zinnen beschrijven een specifieke situatie, en volgen een min of meer vast patroon. Dit is niet het geval voor de zinnen die over het onderwerp van de wet gaan: normen en definities. Deze zinnen beschrijven allerlei verschillende situaties en onderwerpen, en volgen geen vast patroon.

Voor de zinnen die over de wet zelf gaan kan er voor elke standaardzin worden bepaald welke onderdelen daarvan nodig zijn om een model van die zin te maken. Voor een model van een zin die tekst vervangt is bijvoorbeeld de locatie waar de tekst vervangen wordt, de te vervangen tekst en de vervangende tekst nodig. In deze zinnen is de locatie altijd aanwezig als een verwijzing, de tekst die wordt vervangen is gemarkeerd met aanhalingstekens en de vervangende tekst volgt na de (eerste) dubbele punt in de zin. Voor andere zinnen die over de wet gaan, zoals het vastleggen van de citeertitel en het intrekken van een wet, kunnen soortgelijke regels worden gebruikt om de noodzakelijke informatie te verkrijgen. Een handmatige telling binnen een corpus bestaande uit 343 zinnen van deze soort suggereert dat ongeveer 96% van deze zinnen op deze manier kan worden gemodelleerd.

Voor de zinnen die over het onderwerp van de wet gaan bestaan dergelijke standaardzinnen niet. Het is dus ook niet voldoende om slechts voor een aantal standaardzinnen een vast te leggen welke informatie nodig is voor een model. Voor deze zinnen wordt daarom een generiekere aanpak voorgesteld. Voor een normatieve zin wordt de actie of situatie die is toegestaan of verboden beschreven in een frame. Een dergelijk frame bevat informatie zoals de actie, de agens en patiëns. Deze informatie kan uit de zin worden gehaald middels bestaande natuurlijke taalverwerking toepassingen.

Dergelijke frames vormen een (rudimentair) instrument voor het toepassen van de regels. Door een casus in overeenkomstige termen te beschrijven (dus actie, agens, patiëns, etc.) kan

de situatie in de casus vergeleken worden met de situatie in de regels. Indien de situatie overeenkomt, is de regel van toepassing, en is duidelijk of volgens die regel de situatie is toegestaan of niet.

Een soortgelijke methode kan worden toegepast voor conditionele bijzinnen en definities. Als het gaat om een conditionele bijzin en de situatie casus komt overeen met die in de bijzin, dan geeft dit aan dat de regel (hoofdzin) waar de bijzin bij hoort moet worden toegepast. Als het gaat om een definitie, dan geeft een overeenkomst aan dat aan de definitie voldaan doen, en dat regels die van toepassing zijn op het gedefinieerde concept ook van toepassing zijn op deze casus.

Nadat de zinnen zijn gemodelleerd, moeten de losse modellen worden geïntegreerd. Modellen kunnen aan elkaar verbonden worden omdat de zinnen waar ze bij horen naar elkaar verwijzen middels een verwijzing, of doordat ze dezelfde concepten gebruiken. Deze verbanden zijn makkelijk op te sporen. Verwijzingen zijn al opgespoord, en als twee zinnen dezelfde concepten gebruiken, dan worden vaak ook dezelfde woorden gebruikt om dat concept aan te duiden. Echter, zinnen zijn ook vaak verbonden door *common sense* relaties tussen twee concepten die ze gebruiken. Zo zijn bijvoorbeeld de concepten *gift* en *gever* aan elkaar gerelateerd, en als gevolg daarvan zijn ook de zinnen die deze twee verschillende concepten gebruiken aan elkaar gerelateerd. Omdat deze relatie nergens expliciet gedefinieerd wordt in de wet kan dit verband niet automatisch worden gelegd, tenzij we deze kennis toevoegen aan het proces. De hoeveelheid van dergelijke *common sense* informatie is echter enorm, en het is dus onwaarschijnlijk dat alle benodigde informatie beschikbaar is in het automatisch modelleerproces. Dit betekent dat slechts een deel van de integratie automatisch kan worden uitgevoerd.

Aangezien de integratiestap niet automatisch kan worden uitgevoerd, is het niet mogelijk om volledig automatisch modellen van een wettekst te maken. Echter, computerprogramma's kunnen wel helpen om de kennisacquisitie bottleneck te verminderen, aangezien ze succesvol zijn in het bepalen van de structuur van een wettekst, het opsporen van verwijzingen in die tekst, het classificeren van zinnen en het helpen modelleren van die zinnen. Daarnaast hebben ook de tussenresultaten van dit proces hun nut. Zo wordt is het automatisch opsporen van verwijzingen gebruikt bij het bouwen van een semantisch netwerk voor de Belastingdienst, en zijn de patronen die zijn geïdentificeerd voor de zinsclassificatie gebruikt als sjablonen in MetaVex, een tekstverwerkingsprogramma voor wetgeving. Deze toepassingen worden beschreven in hoofdstuk 6.

Om de voorgestelde methode daadwerkelijk in gebruik te nemen, zal hij moeten worden verfijnd. De nauwkeurigheid kan worden verbeterd door meer patronen te verzamelen uit een grotere testset. Ook kunnen een aantal verbeteringen worden doorgevoerd die in dit proefschrift beschreven worden. Daarnaast kunnen wetgevingsjuristen ook helpen bij het verbeteren van de automatische verwerking van wetteksten. Sommige stukken informatie kunnen op een meer precieze manier onder woorden worden gebracht, die ze mogelijk ook beter te begrijpen maakt voor de menselijke lezer.

Niet alleen kan de nauwkeurigheid worden verbeterd; de methode kan ook worden uitgebreid. Met betrekking tot de classificatie zou het wenselijk kunnen zijn om meer subcategorieën te

herkennen. Ook is uitbreiding nodig naast wetten ook andere bronnen aan te kunnen, zoals Koninklijke Besluiten en jurisprudentie. Hiervoor zijn extra patronen nodig, die specifieker voor dergelijke teksten zijn.