Oneindige discussies in één oogopslag

Het handmatig samenvatten van online discussies

Chris Veltman
ANR: 407872
21-08-2014

Bachelorscriptie
Communicatie- en Informatiewetenschappen
Faculteit Geesteswetenschappen
Tilburg University

Begeleider: Dr. S. Wubben
Tweede lezer: Dr. M.M van Zaanen

Abstract

In dit onderzoek is onderzocht hoe mensen een samenvatting van een online discussie maken terwijl ze de informatie uit de discussie classificeren en het relevante selecteren. De online discussies die in dit onderzoek aan bod komen, vinden plaats op een internetforum: een online ruimte waarin mensen met elkaar kunnen discussiëren over bepaalde onderwerpen door het plaatsen van berichten. Het onderwerp van de discussie wordt ook wel het topic genoemd en de gehele discussie op het forum heet een *forumthread*. Er is onderzocht hoe mensen de *forumposts* (reacties in de discussie) classificeren met behulp van een classificatieschema, waarbij één of meerdere vooraf vastgestelde labels toegekend konden worden aan een *forumpost*. Daarnaast is onderzocht hoe relevant mensen de reacties vinden voor de vraag of het onderwerp in een *forumthread*. Tot slot is onderzocht hoe mensen zelf een samenvatting van een *forumthread* schrijven. Hierbij moesten de mensen zelf een samenvatting schrijven waarna deze samenvattingen zijn geanalyseerd en vergeleken. Er hebben 49 respondenten meegewerkt aan dit onderzoek dat is uitgevoerd opdat de resultaten uiteindelijk van hulp kunnen zijn tijdens het creëren van een automatisch samenvattingssysteem voor *forumthreads*.

Er is gebleken dat mensen de meeste reacties als relevant voor het onderwerp beoordelen. Zelfs reacties die ze uiteindelijk niet in hun eigen samenvatting van de *forumthread* verwerken, vinden ze vaak relevant voor het onderwerp. Daarnaast hebben mensen een gemeenschappelijke voorkeur voor reacties die ze in hun samenvatting willen gebruiken. De mate van overeenkomst in het classificeren van de *forumposts* ligt ver uit elkaar. Dit kan komen door het gebruikte classificatieschema, wat aangeeft dat een dergelijk classificatieschema niet werkt bij grote groepen respondenten. Immers, hoe groter de groep, hoe groter de afwijkingen tussen de classificatiekeuzes zullen zijn. Een aanbeveling voor volgende onderzoeken is dan ook om een nieuw classificatieschema op te bouwen specifiek voor grotere groepen respondenten.

De resultaten van dit onderzoek zijn samengevoegd in een dataset die online te vinden is (zie Hoofdstuk 2 Methode).

Inhoudsopgave

Hoofdstuk 1. Inleiding

Momenteel zijn ongeveer drie miljard mensen in de wereld internetgebruikers[1]. Dit aantal groeit elke dag. Deze internetgebruikers zoeken niet alleen informatie op internet, maar voegen zelf ook informatie toe. In de huidige samenleving waarin de technologie continu verbeterd wordt en computers en internet steeds toegankelijker worden, is de informatiestroom zeer groot. In sommige gevallen zelfs zo groot, dat het voor gebruikers bijna niet meer te verwerken is: de hoeveelheid informatie beschikbaar zorgt voor een *information overload* (Eppler & Mengis, 2004). Dit is bijvoorbeeld het geval bij internetfora.

Een internetforum is een online ruimte waarin mensen met elkaar kunnen discussiëren over bepaalde onderwerpen door het plaatsen van berichten. Iemand kan een vraag online zetten, waarop anderen op andere tijdstippen antwoorden en reacties kunnen plaatsen. Op deze manier kunnen er zeer lange discussies ontstaan over zeer verschillende onderwerpen. Het onderwerp van een discussie wordt ook wel een topic of *thread* genoemd. Wanneer iemand een specifieke vraag heeft die al in een internetforum is gesteld, liggen de antwoorden op deze vraag voor het oprapen. Het probleem is echter dat de discussies op internetfora zo lang kunnen worden, dat het voor de lezer alsnog een tijdrovende zoektocht wordt naar een relevante reactie. Niet alle reacties in een internetforum zijn namelijk relevant: veel reacties beantwoorden de vraag niet of zijn zelfs off-topic. Een mogelijke oplossing zou kunnen liggen in het maken van samenvattingen van de internetfora. Door samenvattingen van reacties in een forum op de website te plaatsen, kan de lezer in één korte blik zien wat relevant is. Dit bespaart de lezer een hoop tijd.

Een samenvatting is een aangepaste vorm van een brontekst waarbij de inhoud gereduceerd is door slechts de belangrijke en relevante informatie uit de brontekst te selecteren en in sommige gevallen de inhoud te generaliseren (Jones, 1999). In het geval van een internetforum zou de samenvatting kunnen bestaan uit de meest relevante reacties. Iemand die op zoek is naar het antwoord op de gestelde vraag aan het begin van het forum heeft geen behoefte aan reacties die geen antwoord geven: deze lezer wil dus een gefilterde versie van het internetforum. Welke reacties meer of minder relevant zijn, kan bepaald worden met behulp van een classificatiesysteem: de reacties worden ingedeeld in groepen op basis van overeenkomsten in kenmerken of eigenschappen (Cormack, 1971). Daarna kunnen bepaalde kenmerken gekoppeld worden aan mate van relevantie, waardoor op basis van de klassen duidelijk is welke reacties relevant zijn. Met behulp van classificatiesystemen zouden de meest relevante reacties samengevoegd kunnen worden tot een samenvatting van de discussie. De vraag die in het begin van de forumdiscussie gesteld wordt zou in combinatie met de samenvatting van de meest relevante reacties getoond kunnen worden boven

1 Worldometers, geraadpleegd op 18 maart 2014, van http://www.worldometers.info/nl/

aan de internetdiscussie, zodat de lezer in één oogopslag kan zien wat de belangrijkste reacties zijn.

Hoewel het toevoegen van deze samenvattingen de lezer veel tijd bespaart, kost het de schrijvers zeer veel tijd om voor elk internetforum van alle reacties een samenvatting te maken. Het is niet realistisch om mensen dit handmatig te laten doen. Het zou daarom handiger zijn als deze samenvattingen automatisch gegenereerd zouden kunnen worden. Met de huidige technologische mogelijkheden wordt het steeds gebruikelijker teksten automatisch te genereren in plaats van deze handmatig te schrijven. Vooral naar het automatisch genereren van samenvattingen is al veel onderzoek gedaan (Jones, 1999; McKeown, Klavans, Hatzivassiloglou, Barzilay & Eskin 1999; Knight & Marcu, 2002).

Om een computer te trainen samenvattingen te genereren van bijvoorbeeld reacties in een forum, moet eerst onderzocht worden welke keuzes mensen maken tijdens het maken van een samenvatting. Welke reacties vinden mensen bijvoorbeeld meer of minder relevant? Daarnaast is het ook van belang om te weten of alle mensen hierin dezelfde keuzes maken, of dat sommige mensen bepaalde reacties relevanter vinden dan anderen.

De onderzoeksvraag luidt daarom: Hoe maken mensen een samenvatting van een online discussie terwijl ze de informatie uit de discussie classificeren en het relevante selecteren?

In dit onderzoek zal specifiek gekeken worden naar het Viva forum, "het populaire forum voor vrouwen over onderwerpen zoals eten, gezondheid, mode, seks, werk en wonen." [2] Daarnaast zullen slechts de topics met vragen als beginpositie gebruikt worden. Dit zijn dus de *forumthreads* waarin een vraag wordt besproken en mogelijke antwoorden worden gezocht, in tegenstelling tot *forumthreads* waar slechts een discussie gaande is over bijvoorbeeld een mening. De topics die in dit onderzoek gebruikt worden zijn dus op zoek naar een specifiek antwoord op de vraag.

Dataset

In dit onderzoek zal een dataset worden opgebouwd omtrent drie *forumthreads*. De dataset zal de volgende gegevens bevatten: de relevantiebeoordelingen van de *forumposts* uit de drie *forumthreads*, de classificering van de *forumposts* en de handmatig gemaakte samenvattingen van de drie *forumthreads* van een groep respondenten.

2 Viva, *Viva Forum.*, geraadpleegd op 5 maart 2014, van http://forum.viva.nl/

Hoofdstuk 2. Theoretisch Kader

2.1 Handmatig samenvatten

Volgens Jones (1999) is een samenvatting een aangepaste vorm van een brontekst waarbij de inhoud gereduceerd is door slechts de belangrijke en relevante informatie uit de brontekst te selecteren en in sommige gevallen de inhoud te generaliseren. Het samenvatten van een tekst of verschillende teksten is een moeilijke taak (Jones, 1999). Om te kunnen samenvatten moet iemand begrijpen wat belangrijk is in een tekst en hier aandacht voor hebben ten koste van onbeduidende informatie (Brown, 1983). Dit lijkt een vaardigheid die mensen ontwikkelen naarmate ze ouder worden: in het onderzoek van Brown, Day en Jones (1983) is gebleken dat oudere middelbare school leerlingen beter presteren dan jongere leerlingen in hun gevoeligheid voor wat belangrijk is in de tekst. De vaardigheid om informatie te verwerken vereist namelijk beoordelingsvermogen, inzet, kennis en strategieën. Op jongere leeftijd passen kinderen bij het samenvatten slechts simpele schrapprocedures toe (Brown, 1983).

Een samenvatting wordt gemaakt aan de hand van een brontekst (Jones, 1999). Een brontekst heeft drie eigenschappen, namelijk linguïstisch, domein en communicatief (Barzilay & Elhadad, 1997). Elk van deze drie tekstaspecten kan als basis voor de samenvatting gekozen worden. Wanneer een samenvatting aan de hand van het linguïstische aspect van een brontekst wordt gemaakt, wordt er slechts gekeken naar het soort woorden wat gebruikt is. In dit geval wordt de samenvatting bijvoorbeeld gecreëerd met woorden die het meest in de brontekst voorkomen, met de onderliggende gedachte dat de woorden die het meeste voorkomen ook de belangrijkste concepten van de tekst weergeven (Barzilay & Elhadad, 1997). Wanneer een samenvatting aan de hand van het communicatieve aspect van een brontekst wordt gemaakt, kan de samenvatting worden opgebouwd op basis van een semantische analyse van de brontekst. In dit geval wordt er dus gekeken naar de betekenissen van de woorden en hoe er stukken tekst geschrapt kunnen worden, zonder dat de tekst de coherentie verliest. Wanneer er rekening wordt gehouden met het linguïstische aspect is de coherentie van de samenvatting niet zo van belang. In het geval van het domein aspect van een brontekst, wordt er tijdens het samenvatten rekening gehouden met het domein van de tekst. Het domein kan variëren van een krantenartikel tot een wetsdocument (Aone, Okurowski, Gorlinsky & Larsen, 1999). Er wordt in dit geval tijdens het samenvatten gebruik gemaakt van algemeen beschikbare bronnen die informatie bevatten over het domein van de brontekst, waardoor het niet nodig is dat de brontekst volledig begrepen wordt. Volgens Barzilay en Elhadad (1997) is het belangrijk dat er met de aspecten rekening gehouden wordt, omdat de tekst in een samenvatting anders te ver versimpeld wordt, waardoor de kwaliteit van de samenvatting omlaag gaat. De samenvatting kan bijvoorbeeld minder leesbaar worden of de boodschap wordt niet

meer op dezelfde manier overgebracht. Het is dus ook voor mensen noodzakelijk om rekening te houden met de aspecten van een brontekst om een goede samenvatting te kunnen maken. Het linguïstische aspect lijkt het belangrijkste omdat deze er voor zorgt dat een tekst correct en leesbaar is.

Goldstein et al. (1999) beschrijven een samenvatting als een algemene tekst die korter is dan de brontekst en die de punten van de tekst reflecteert die de schrijver van de samenvatting belangrijk acht. Samenvattingen van de menselijke hand zijn dus vrij subjectief, aangezien de suggestie wordt gewekt dat niet iedereen hetzelfde belangrijk vindt in een tekst. Iemand die een samenvatting maakt, wil volgens Goldstein et al. dat deze zo algemeen mogelijk is en voegt dus alleen de punten toe die hij of zij zelf belangrijk acht. Daarnaast gebruikt niet iedereen dezelfde woorden in de samenvatting en hanteert niet iedereen dezelfde schrijfstijl, waardoor handmatig geschreven samenvattingen nog meer van elkaar verschillen (Copeck & Szpakowicz, 2004). Ook bestaan er verschillende samenvattingsstrategieën. In de volgende paragraaf worden enkele van deze strategieën beschreven.

2.2 Samenvattingsstrategieën

Een onderscheid dat veel wordt gemaakt bij het samenvatten van een tekst, is die tussen extractief en abstractief samenvatten. Bij extractief samenvatten worden de belangrijkste zinnen en paragrafen geselecteerd, waarna in deze zinnen en paragrafen geschrapt wordt opdat de lengte van de nieuwe tekst korter wordt. Hierdoor komt de samenvatting qua woordgebruik zeer overeen met de brontekst. In het geval van abstractief samenvatten worden de zinnen en woorden uit de brontekst ook anders verwoord en gecombineerd, waardoor er compleet nieuwe zinnen ontstaan (Carenini & Cheung, 2008). In het geval van abstractief samenvatten moet de schrijver de brontekst dus ook begrijpen en zelf op een andere manier en met minder woorden samenvatten (Gupta & Lehal, 2010).

De strategieën die mensen aanhouden tijdens het maken van een samenvatting veranderen naarmate er meer ervaring is met samenvatten en deze vaardigheid meer ontwikkelt. Zo beschrijven Brown, Day en Jones (1983) de *copy-delete* strategie (kopieer-schrap strategie). Deze strategie bevat drie stappen: eerst worden de tekstelementen opeenvolgend gelezen, waarna besloten wordt welke elementen toegevoegd of geschrapt moeten worden en tot slot, als het tekstelement moet worden toegevoegd, of het min of meer gekopieerd moet worden of met minder woorden dan in de tekst in de samenvatting verwerkt moet worden. Deze strategie komt overeen met het extractief samenvatten en wordt vooral gevolgd door jongere kinderen die nog niet de noodzaak voelen om de hoofdpunten van een tekst in hun eigen woorden te noteren.

Wanneer de kinderen ouder worden en gaan studeren, gebruiken ze tijdens het maken van

een samenvatting een radicaal andere strategie dan de *copy-delete* strategie (Brown et al, 1983). Ze hebben dan meer ervaring met het samenvatten van teksten en hebben de behoefte om de samenvatting in hun eigen woorden te zetten. Tijdens het samenvatten wijken de studenten af van zowel de gebruikte woorden als de chronologische volgorde in de tekst. De studenten combineren informatie uit verschillende paragrafen, delen zelf de volgorde van onderwerpen in en noteren de samenvatting in hun eigen woorden. Deze techniek komt overeen met het abstractieve samenvatten.

Volgens Nenkova en Vanderwende (2005) verwerken mensen sneller woorden in hun samenvatting wanneer deze woorden vaak herhaald zijn in de brontekst. Toch is er geen daadwerkelijke samenvattingsstrategie wat betreft de frequentie van woorden in de brontekst, aangezien woorden die relatief weinig voorkomen, ook in de samenvattingen worden gebruikt. Bovendien gebruiken mensen tijdens het maken van een samenvatting ook woorden die niet in de brontekst voorkomen (Copeck & Szpakowicz, 2004).

2.3 Automatisch samenvatten

Zoals in de vorige paragraaf al is genoemd, is een samenvatting volgens Jones (1999) een aangepaste vorm van een brontekst waarbij de inhoud gereduceerd is door slechts de belangrijke en relevante informatie uit de brontekst te selecteren en in sommige gevallen de inhoud te generaliseren. Een samenvatting, of abstract, is een tekst met een kleinere lengte dan de brontekst waar alleen de hoofdpunten in staan vermeld. Jones (1999) benadrukt dat het maken van een samenvatting een moeilijke taak is. Tijdens het samenvatten moet de belangrijke inhoud uit een tekst gehaald worden, waarbij de inhoud zowel informatie als de uitdrukking ervan kan zijn, en het belangrijke zowel dat wat essentieel is en dat wat saillant is (Jones, 1999). Doordat het maken van een samenvatting voor een mens al een moeilijke taak is, is het van belang dat bij het trainen van een computer met de juiste dingen rekening wordt gehouden. Zo beschrijft Jones (1999) in haar onderzoek dat op dit moment niet verwacht moet worden dat de programma's die automatisch samenvatten kunnen wedijveren met het handmatige samenvatten. Tijdens het creëren en trainen van dit soort programma's moet dit dus ook niet nagestreefd worden. Mensen kunnen tijdens het maken van samenvattingen immers keuzes maken over wat relevant is en wat niet en de relevante informatie op een logische en creatieve wijze samenvoegen in een samenvatting. Computers kunnen weliswaar ook de relevante informatie uit een tekst halen, maar deze slechts samenvatten op een grammaticaal correcte manier. Ook volgens Goldstein, Kantrowitz, Mittal en Carbonell (1999) is het moeilijk om de kwaliteit van een menselijke samenvatting na te bootsen, aangezien er teveel variatie is in schrijfstijlen, genres, syntactische constructies, lexicale items etc. Computers kunnen de informatie in een tekst niet automatisch koppelen aan gerelateerde kennis over de wereld die niet in de tekst voorkomt: een computer kan niet redeneren. Daarom is het momenteel ook moeilijk om

een computer abstractief te laten samenvatten want hiervoor is begrip nodig van meer dan slechts dat wat in de tekst staat. Hierdoor gebruiken mensen tijdens het maken van een samenvatting ook woorden die niet in de brontekst voorkomen (Copeck & Szpakowicz, 2004) en die een samenvattingssysteem niet zelf kan toevoegen.

2.4 Samenvattingssysteem

Tijdens het maken van een samenvatting is er, zoals in bovenstaande paragraaf is uitgelegd, een onderscheid te maken tussen extractief en abstractief samenvatten. Bij extractief samenvatten worden de belangrijkste zinnen en paragrafen geselecteerd, waarna in deze zinnen en paragrafen geschrapt wordt opdat de lengte van de nieuwe tekst korter wordt. In het geval van abstractief samenvatten worden er nieuwe zinnen gegenereerd aan de hand van informatie uit de brontekst (Carenini & Cheung, 2008). Wanneer er bij het samenvatten van een online discussie voor wordt gekozen om de relevante reacties op te sommen, is er dus sprake van extractief samenvatten. Er worden immers delen van de brontekst overgenomen in de samenvatting zonder dat de tekst wordt geherformuleerd. Carenini en Cheung (2008) noemen dit voorbeeld ook in hun onderzoek naar de effecten van beide samenvattingsstrategieën tijdens het samenvatten van documenten die meningen en voorkeuren bevatten, zoals het geval is in online discussies. Om een abstractieve samenvatting automatisch te genereren moet de computer in staat zijn de natuurlijke taal goed te kunnen verwerken (Gupta & Lehal, 2010). Zo moet de computer de semantische representatie en de inferenties van de brontekst kunnen verwerken. Daarnaast moet de computer zelf natuurlijke taal kunnen genereren. Deze vaardigheden staan voor computers momenteel nog in een beginfase waardoor het op dit moment gebruikelijker is gebruik te maken van extractieve samenvattingssystemen (Jackson & Moulinier, 2007).

Systemen die gemaakt zijn om extractief samen te vatten identificeren meestal eerst de zinnen in de brontekst die het meest van belang zijn voor het algehele begrip van de brontekst (Alguliey & Aliguliyev, 2009). In de meeste gevallen maken deze samenvattingssystemen gebruik van een metriek op basis van overeenkomsten of centraliteit om de belangrijke zinnen te identificeren. Zo kan er bijvoorbeeld een samenvatting worden geproduceerd die gebaseerd is op links gemaakt tussen paragrafen in de brontekst: in dit geval wordt er gebruik gemaakt van een map/model die de relaties tussen stukken tekst weergeeft waardoor duidelijk is dat de gelinkte stukken tekst semantisch gerelateerd zijn aan elkaar (Alguliey & Aliguliyev, 2009). Ook kan er gebruik worden gemaakt van clustering: het vinden van natuurlijke groepen of clusters en het identificeren van interessante patronen in de brontekst op basis van bepaalde overeenkomsten (Alguliey & Aliguliyev). Extractieve samenvattingssystemen zoeken in sommige gevallen dus naar overeenkomsten tussen zinnen in de gehele tekst om daardoor te kunnen identificeren welke zinnen

van belang zijn voor het begrip van de brontekst.

Een samenvattingssysteem moet meer kunnen dan slechts vaststellen welke stukken tekst in een brontekst belangrijk zijn (Knight & Marcu, 2002). De samenvatting moet namelijk uiteindelijk een samenhangend geheel vormen. Een programma moet niet slechts de belangrijke feiten uit een tekst opsommen, maar hier ook een leesbaar en grammaticaal correct geheel van maken waardoor de lezer daadwerkelijk tijd bespaart door het lezen van de samenvatting. Volgens Knight en Marcu (2002) kan deze taak vergeleken worden met die van zinscompressie. Bij zinscompressie moet ook rekening gehouden worden met alle belangrijke informatie en hoe deze informatie grammaticaal correct uitgedrukt kan worden, alleen op een kleinere schaal. Volgens Cohn en Lapata (2008) kan automatische zinscompressie ook wel beschreven worden als het creëren van een grammaticale samenvatting van een enkele zin waarbij zo min mogelijk informatie verloren gaat. In het geval van automatische zinscompressie in verhouding tot het samenvatten van teksten is het belangrijk dat er rekening gehouden wordt met het feit dat de tekst niet slechts korter gemaakt moet worden, maar dat ook enkel dat wat relevant is gebruikt wordt. Slechts bij de relevante stukken tekst moet automatische zinscompressie worden toegepast.

Volgens McKeown et al. (1999) is het bovendien van belang dat de informatie in de samenvatting niet dubbelop moet zijn: als er teveel gelijksoortige informatie wordt toegevoegd, wordt de samenvatting langer dan noodzakelijk waardoor de lezer alsnog veel tijd verliest aan het lezen van overbodige informatie.

Het is zeer moeilijk om een standaard voor algemene samenvattingen te maken, aangezien verschillende mensen ook verschillende samenvattingen maken van eenzelfde brontekst. Daarom moet bij het maken van een samenvattingssyteem niet verwacht worden dat er één perfect model gebruikt kan worden. Daarnaast is het van belang dat de samenvattingssystemen geëvalueerd worden: voldoet het samenvattingssyteem wel aan de eisen? Volgens Carenini en Cheung (2008) zijn er verschillende manieren om samenvattingssystemen te evalueren, waaronder het vragen van menselijke beoordelingen, taakgerichte benaderingen en automatische methodes. Bij menselijke beoordelingen moeten mensen de kwaliteit van de samenvatting beoordelen. Een taakgerichte methode test de effectiviteit van een samenvattingssysteem voor het bedoelde doel (McKeown, Passonneau, Elson, Nenkova & Hirschberg, 2005). Een voorbeeld van een automatische methode is ROUGE (Lin, 2004, in: Carenini & Cheung, 2008). Deze methode geeft een score op basis van overeenkomsten in de volgorde van woorden in de handmatig geschreven modelsamenvatting en de samenvatting gegenereerd door de computer. Hoewel de evaluaties die ROUGE geeft vaak vrij goed overeen komen met hoe mensen de samenvattingen beoordelen, geeft de methode geen inzicht in de specifieke zwaktes en krachten van de samenvatting. Er wordt bij deze methode namelijk slechts gekeken naar de overeenkomst in woorden, niet naar aspecten zoals zinsopbouw, waardoor een

samenvatting met een hoge evaluatie alsnog onleesbaar kan zijn.

2.5 Soorten samenvattingssystemen

Het wereldwijde web blijft continu groeien waardoor de vraag naar automatische samenvattingssystemen steeds groter wordt. Vooral de grote hoeveelheid nieuws en informatie online heeft er voor gezorgd dat er verschillende multi-document samenvattingssystemen zijn geproduceerd (Nenkova & Vanderwende, 2005). Een multi-document samenvattingssysteem is een systeem dat één samenvatting produceert van verschillende, individuele documenten. Automatische samenvattingssystemen kunnen de nieuwe informatie continu automatisch verwerken. Het grootste probleem met de samenvattingssystemen is het selecteren van de inhoud: beslissen welke zinnen uit de documenten en teksten belangrijk genoeg zijn om in de samenvatting te worden toegevoegd (Nenkova & Vanderwende, 2005). Zoals eerder aangegeven, is het maken van een samenvatting subjectief, aangezien niet iedereen hetzelfde relevant lijkt te vinden (Goldstein et al., 1999). Voor de mens is het maken van een algemene samenvatting al een moeilijke taak, waardoor het nog moeilijker is om het een computer aan te leren.

In het onderzoek van Carenini en Cheung (2008) komt zowel een abstractief als extractief multi-document samenvattingssysteem aan de orde. Beide systemen zijn specifiek ontwikkeld voor het evaluatieve domein. Hier valt een internetdiscussie ook onder, aangezien er onder andere meningen worden verkondigd. Als eerste moeten de systemen de zinnen met meningen in de brontekst identificeren. Hierbij moet rekening gehouden worden met de eigenschappen van datgene wat geëvalueerd wordt, de kracht van de mening en of de evaluatie ofwel positief ofwel negatief is. Bij eigenschappen van datgene wat geëvalueerd wordt, moet gedacht worden aan waar de reactie toepassing op heeft. In het onderzoek van Carenini en Cheung (2008) wordt het voorbeeld gegeven van een reactie uit een corpus van klanten reviews over een camera, die aangeeft dat de foto kwaliteit uitstekend is. Deze reactie geeft dus een mening over de eigenschap fotokwaliteit. Daarnaast is deze reactie zeer positief.

Het abstractieve samenvattingssysteem in het onderzoek van Carenini en Cheung (2008) is de *Summarizer of Evaluative Arguments* (Samenvattingssysteem van Evaluatieve Argumenten, SEA). In dit systeem wordt de inhoud van een tekst gecategoriseerd op basis van de eigenschappen van datgene wat geëvalueerd wordt, die in de reacties voorkomen. Het belang van de eigenschappen wordt gebaseerd op de hoeveelheid evaluaties met deze eigenschappen en de kracht van de evaluaties. De meeste belangrijke eigenschappen komen in de uiteindelijke samenvatting met één zin, gegenereerd uit een patroon gebaseerd op de hoeveelheid en de verspreiding van de polariteit (positief of negatief) en kracht van de evaluaties van de eigenschappen.

Het extractieve samenvattingssysteem dat genoemd wordt in het onderzoek van Carenini en

Cheung (2008) heet MEAD. MEAD categoriseert de eigenschappen op basis van de hoeveelheid van zinnen die een evaluatie geven over die eigenschap. Daarna wordt voor elke eigenschap een zin geselecteerd totdat de woordlimiet is bereikt. De zin die voor elke eigenschap wordt geselecteerd, is de zin met de hoogste som van polariteit en kracht evaluaties voor een eigenschap. Hierdoor worden vooral de zinnen geselecteerd die meerdere eigenschappen benoemen. De zinnen worden vervolgens geordend met behulp van een hiërarchie waardoor de meer abstracte eigenschappen eerder genoemd worden dan de meer specifieke eigenschappen.

Een andere methode waarmee een samenvattingssysteem kan werken is zinscompressie. Zinscompressie is, zoals eerder genoemd, het samenvatten van een enkele zin opdat de zin korter wordt, waarbij de zin nog steeds grammaticaal correct is en de belangrijkste informatie van de bronzin bevat (Knight & Marcu, 2002). Zinscompressie komt overeen met het samenvatten van een brontekst, alleen is het op een kleiner niveau. Een voorbeeld van een zinscompressie methode is het *noisy channel* (ruis kanaal) model (Knight & Marcu, 2002). In dit model wordt er naar een lange zin gekeken en wordt er vanuit gegaan dat dit origineel een korte zin was, waarna iemand er extra, optionele tekst aan heeft toegevoegd. Compressie is dan een kwestie van het identificeren van de originele korte zin. De extra, optionele tekst wordt gezien als *noise,* ofwel "ruis".

Het compressie algoritme beschreven in het onderzoek van Knight en Marcu (2002) wordt geëvalueerd met behulp van een corpus met handmatig geschreven zinnen. Uit dit corpus worden aselect zinnen gekozen die vervolgens naast de zinnen worden gelegd die gegenereerd zijn door het algoritme. Hierna werden deze zinnen voorgelegd aan mensen die in de veronderstelling waren dat alle zinnen automatisch gegenereerd waren. Deze mensen moesten beoordelen hoe goed de systemen de meest belangrijke woorden uit de originele zin hadden geselecteerd op een schaal van 1 tot 5.

Er zijn verschillende manieren om een programma automatisch een samenvatting te laten maken. Ten eerste moet er een selectie worden gemaakt van een klein aantal betekenisvolle zinnen uit de brontekst (Teufel & Moens, 1997). Deze selectie komt enigszins overeen met het selecteren van relevante reacties in een topicthread. In het onderzoek van Kupiec, Pedersen en Chen (1995, in: Teufel & Moens, 1997) blijkt dat dit een classificatie taak is. Op basis van een corpus met technische papers met handmatig geschreven samenvattingen, waarbij de technische papers de bronteksten zijn en de samenvattingen de doelteksten, identificeert hun systeem de zinnen in de tekst die ook in de samenvatting voorkomen, waarna een model wordt gecreëerd waarin duidelijk wordt welke zinnen waardevol genoeg zijn om in een samenvatting voor te komen. Op basis van een corpus is het dus mogelijk een systeem aan te leren welke informatie uit een brontekst relevant is. Een dergelijke dataset bevat een brontekst en een bepaald aantal handmatig gemaakte samenvattingen van die brontekst. Aan de hand van deze dataset kan de computer vervolgens een

modelsamenvatting maken.

Volgens Goldstein et al. (1999) zijn mensen in veel situaties geïnteresseerd in andere feiten dan die in de algemene samenvatting, waardoor er behoefte is aan 'vraag-relevante' samenvattingen: samenvattingen die precies dat laten zien waar op dat moment door de lezer vraag naar is. Dit komt overeen met de behoefte van de lezer van een online discussie: de lezer heeft een vraag of gedachte en wil weten wat anderen hierover denken.

Om een automatisch samenvattingssysteem te creëren moet gebruik gemaakt worden van handmatig gemaakte samenvattingen (Copeck & Szpakowicz, 2004). Zoals in het eerder genoemde onderzoek van Kupiec et al. (1995, in: Teufel & Moens, 1997) kan een corpus met bronteksten en handmatig geschreven samenvattingen aangemaakt worden. De samenvattingen kunnen een heuristiek suggereren tijdens het systeemontwerp, als trainingsdata functioneren en optreden als een standaard waartegen de automatisch gegenereerde samenvattingen geëvalueerd kunnen worden (Copeck & Szpakowicz). Door gebruik te maken van handmatig gemaakte samenvattingen om samenvattingssystemen te trainen, leren de systemen welke factoren voor mensen van belang zijn tijdens het maken van een samenvatting. Deze factoren kunnen bijvoorbeeld classificatietaken zijn, die in de volgende paragraaf beschreven zullen worden. Hierna kan het systeem de factoren op dezelfde manier toepassen tijdens het genereren van een samenvatting.

2.6 Automatische classificatie

Een classificatie plaatst entiteiten in vooraf gedefinieerde klassen, waardoor alle entiteiten in een klasse op een bepaalde manier met elkaar verbonden zijn (Cormack, 1971). Het is een manier om een onderscheid te maken in de gegeven informatie. Een voorbeeld van een classificatiemethode is de *k-nearest neighbors* (naaste buren) methode (Cover & Hart, 1967; Weinberger, Blitzer & Saul, 2006). De *k-nearest neighbors* methode classificeert elk ongedefinieerd voorbeeld met het label dat de meerderheid van de dichtstbijzijnde "buren" in de training set heeft. Een zin of woord krijgt dus de classificatie die de meerderheid van de zinnen of woorden heeft, die in de buurt staan. Hierbij wordt gebruik gemaakt van een metriek die de dichtstbijzijnde buren identificeert.

Nenkova en Vanderwende (2005) geven aan dat het gebruikelijk is om zinnen uit de brontekst te classificeren waardoor duidelijk wordt welke zinnen belangrijk en relevant genoeg zijn om aan de samenvattingen te mogen worden toegevoegd. Dit kan gebeuren door een bepaald belang aan de zinnen toe te wijzen op basis van verschillende eigenschappen (Schiffman, Nenkova & McKeown, 2002). Voor deze eigenschappen moet dan heuristiek besloten zijn in hoeverre ze belangrijk zijn. Hierna kunnen de zinnen die het hoogst scoren geselecteerd worden voor de samenvatting. Voorbeelden van dergelijke eigenschappen zijn woordfrequentie en de mate van overeenkomst. Zo kan een zin in de brontekst een hoge classificatie krijgen wanneer er woorden in

voorkomen die met een onderwerp te maken hebben dat veel terugkomt in de algehele tekst.. Hierbij wordt gebruik gemaakt van associaties: gelijkwaardige woorden als 'restaurant' en 'ober' vallen onder hetzelfde onderwerp (Schiffman et al., 2002). Andere eigenschappen waar tijdens het classificeren van zinnen rekening mee gehouden zou kunnen worden, is de aanwezigheid van woorden die tevens in de titel van de tekst of tussenkopjes voorkomen. De woorden die in de titel van de tekst of in de titels van paragrafen staan, worden vaak gezien als relevante woorden voor het onderwerp van de tekst (Shiffman et al., 2002).

Een classificatieschema specifiek voor de berichten in webforums is te vinden in artikel van Kim, Wang en Baldwin (2010). Dit schema maakt gebruik van het zogenaamde *Dialogue Act Tagging* waarmee de functie van een uiting of bericht ten opzichte van de gehele dialoog wordt gebruikt. De berichten worden dus geclassificeerd op basis van hun functie in het gesprek. Daarnaast kunnen met dit schema vraag-antwoord paren geïdentificeerd worden (bijvoorbeeld berichten die antwoord geven op een vraag in een eerder bericht), wat toepasselijk is op *forumthreads*, waarin verschillende losse berichten staan die wel op elkaar kunnen aansluiten. De labelset uit het classificatieschema van Kim, Wang en Baldwin (2010) bevat twaalf categorieën, zoals te zien is in Schema 1. Er zijn twee hoofdcategorieën, namelijk "vraag" en "antwoord", en drie individuele klassen, namelijk "resolutie", "reproductie" en "anders". De hoofdklasse "vraag" is opgedeeld in vier sub-klassen, namelijk "vraag" (deze klasse is bedoeld voor de eerste post in een forum, met als het ware de hoofdvraag), "toevoeging", "bevestiging" en "correctie". De hoofdklasse "antwoord" is daarnaast opgedeeld in vijf sub-klassen, namelijk "antwoord", "toevoeging", "bevestiging", "correctie" en "bezwaar". Dit classificatieschema lijkt dus geschikt voor het classificeren van berichten in een *forumthread*.

Schema 1

Classificatieschema; de Klassen en Bijbehorende Beschrijvingen

Klasse	**Beschrijving**
VRAAG-VRAAG	De post bevat een nieuwe vraag. Dit label is gereserveerd voor de eerste post in een thread
VRAAG-TOEVOEGING	De post voegt iets toe aan de vraag door extra informatie te geven of een nieuwe vraag te stellen die doorgaat op de eerste vraag
VRAAG-BEVESTIGING	De post wijst op fout(en) in een vraag zonder deze fouten te verbeteren, of de post bevestigt bepaalde details van de vraag
VRAAG-CORRECTIE	De post corrigeert fout(en) in een vraag
ANTWOORD-ANTWOORD	De post geeft een mogelijk antwoord op de vraag.
ANTWOORD-TOEVOEGING	De post vult een antwoord aan door extra informatie te geven
ANTWOORD-BEVESTIGING	De post benoemt fout(en) in een antwoord zonder deze fout(en) te verbeteren, of de post bevestigt details van een antwoord.
ANTWOORD-CORRECTIE	De post verbetert fout(en) in een antwoord.
ANTWOORD-BEZWAAR	De post maakt bezwaar tegen een antwoord op basis van ervaring of theorie.
RESOLUTIE	De post bevestigt dat een antwoord werkt op basis van ervaring.
REPRODUCTIE	De post ofwel (1) bevestigt dat hetzelfde probleem ervaren wordt, of (2) bevestigt dat een antwoord zou moeten werken.
ANDERS	De post behoort niet tot één van bovenstaande klassen.

2.7 Subvragen

Door de huidige informatiestroom is er een grote vraag naar automatische samenvattingssystemen. Op een forum zoals het Viva forum wordt continu gereageerd en nieuwe informatie toegevoegd, waardoor een programma wat van deze reacties automatisch een samenvatting kan maken van toegevoegde waarde is. Voordat een computer hiervoor getraind kan worden, moet onderzocht worden hoe mensen een samenvatting zouden maken van het forum. Ten eerste is het van belang te onderzoeken of mensen overeenkomen in het classificeren van *forumposts*. Daarom is de volgende subvraag opgesteld:

Subvraag 1: In hoeverre komen mensen overeen in het classificeren van *forumposts* uit een *forumthread?*

Daarnaast is het voor het schrijven van samenvattingen vooral belangrijk om erachter te komen welke reacties mensen in een forum belangrijk en relevant vinden.

Subvraag 2: In hoeverre kunnen reacties op vragen gesteld op internetfora door mensen worden beoordeeld als relevant of niet relevant?

Het is ook van belang te onderzoeken hoe mensen de reacties vervolgens in een samenvatting zouden verwerken. Daarnaast moet er onderzocht worden of er overeenkomsten te vinden zijn in de manier waarop mensen samenvatten, aangezien een samenvattingssysteem een bepaalde manier van samenvatten geleerd moet worden. Hiervoor is de volgende subvraag opgesteld:

Subvraag 3: In hoeverre komen door mensen gemaakte samenvattingen aan de hand van relevante reacties op vragen op internetfora overeen?

Hoofdstuk 3. Methodologie

Er is onderzoek gedaan naar de manier waarop mensen forumthreads samenvatten. Hierbij is gebruik gemaakt van een vragenlijst waarin verschillende onderdelen aan bod kwamen. Deze vragenlijst, de manier waarop de respondenten zijn geselecteerd en hoe de procedure in werking is gegaan, staan beschreven in dit hoofdstuk.

3.1 Instrumentatie

Voor deze studie is gebruik gemaakt van een vragenlijst. Deze vragenlijst is gemaakt met behulp van Qualtrics: een online survey-software waarmee surveys ontworpen en afgenomen kunnen worden. Er is gekozen voor Qualtrics omdat deze software langere surveys met uitgebreidere vragen en antwoordopties kan verwerken, waarvan in dit onderzoek sprake is.

Figuur 1: Forumbericht zoals gepresenteerd op het Viva Forum

In de vragenlijst zijn voorbeelden van forumberichten met reacties toegevoegd (Figuur 1). Deze

forumberichten zijn overgenomen van bestaande *forumthreads* op de Viva-website[3]. In de vragenlijst stonden slechts de daadwerkelijke forumberichten en reacties, en daarbij de gebruikersnaam van de schrijver (Bijlage 1). De profielfoto's van de schrijvers en de datum en de tijd van het plaatsen van de berichten zijn dus niet meegenomen, omdat deze geen verdere toevoeging hadden aan het onderzoek. De gebruikersnamen daarentegen maken aan de lezer duidelijk wanneer verschillende reacties door dezelfde schrijver zijn geschreven. Er zijn in totaal drie verschillende *forumthreads* in de vragenlijst toegevoegd, met elk in totaal acht *posts* (waarvan één het het eerste bericht die de *forumthread* initieert en zeven *posts* met reacties). Er is voor dit aantal gekozen zodat de vragenlijst niet te lang werd, aangezien er in het lezen en begrijpen van het classificatieschema al veel tijd ging zitten, en er alsnog een goed aantal respondenten behaald kon worden. Daarnaast bevatten alle *forumthreads* een eerste bericht met een vraag, waardoor er in de reacties een antwoord werd verwacht. Hierdoor had de *forumthread* een specifiek doel, in plaats van dat er slechts over een onderwerp gepraat of gediscussieerd wordt.

De respondenten werd gevraagd de reacties in de *forumthreads* te classificeren met behulp van het classificatieschema van Kim, Wang en Baldwin (2010), dat ook genoemd is in hoofdstuk 1 paragraaf 5, automatische classificatie (Bijlage 2). Dit classificatieschema stond boven elk van de drie *forumthreads*. Hierdoor kon de respondent tijdens het invullen van de vragen terugkijken naar het schema. In dit schema wordt er ten eerste een onderscheid gemaakt in twee hoofdklassen, namelijk de *posts* met een vraag en die met een antwoord. De inhoud van beide hoofdklassen wordt daarna specifieker geclassificeerd in subklassen. De hoofdklasse met *posts* met een vraag is opgedeeld in vier subklassen, namelijk vraag, toevoeging, bevestiging en correctie.

De vraag-vraag klasse bevat *posts* met een nieuwe vraag, dit label is bedoeld voor de eerste post van een thread. Dan de vraag-toevoeging klasse, die *posts* bevat die iets toevoegen aan de vraag door extra informatie te geven of een nieuwe vraag te stellen die doorgaat op de eerste vraag. Ten derde de vraag-bevestiging klasse die *posts* bevat die wijzen op fouten in een vraag zonder deze fouten te verbeteren, of bepaalde details van de vraag bevestigen. De vraag-correctie klasse beslaat de *posts* die fout(en) in een vraag corrigeren. Naast deze subklassen is ook de hoofdklasse met *posts* met een antwoord opgedeeld in vijf subklassen, namelijk antwoord, toevoeging, bevestiging, correctie en bezwaar. De eerste subklasse, namelijk de antwoord-antwoord klasse, bevat de *posts* die een mogelijk antwoord op de vraag geven. De antwoord-toevoeging klasse bevat de *posts* die een antwoord aanvullen door extra informatie te geven. Dan de antwoord-bevestiging klasse, die de posts bevat die fout(en) in een antwoord benoemen zonder deze fouten te verbeteren, of details van een antwoord bevestigen. De antwoord-correctie klasse bevat de *posts* die fout(en) in een antwoord verbeteren. Tot slot de antwoord-bezwaar klasse waarin de *posts* staan die een bezwaar maken tegen

3 Viva, *Viva Forum.*, geraadpleegd op 5 maart 2014, van http://forum.viva.nl/

een antwoord op basis van ervaring of theorie.

Naast de twee hoofdklassen die ofwel een vraag of antwoord bevatten, of hier betrekking op hebben, zijn er nog drie overige klassen. Ten eerste de resolutie-klasse, die *posts* bevat die bevestigen dat een antwoord werkt op basis van implementatie. Daarnaast de reproductie-klasse, die *posts* bevat die ofwel bevestigen dat hetzelfde probleem ervaren wordt of bevestigen dat een antwoord zou moeten werken. Tot slot de anders-klasse, die de overige *posts* bevat die niet in één van de klassen te plaatsen zijn.

Ook moesten de respondenten de reacties beoordelen op relevantie met behulp van een vijfpuntsschaal. Bij elke *forumpost* stond de volgende zin: "ik vind deze post relevant". Antwoord één stond voor "eens", antwoord twee voor "enigszins eens", antwoord drie voor "neutraal", antwoord vier voor "enigszins oneens" en antwoord vijf voor "oneens". Tot slot moesten de respondenten van elk van de drie *forumthreads* een samenvatting schrijven. Hierbij werd er geen minimum of maximum aan woorden gehanteerd: de respondenten waren geheel vrij om een samenvatting te schrijven die naar hun eigen mening handig en correct was.

3.2 Procedure

De respondenten zijn zowel door middel van sociale media als persoonlijk benaderd. De respondenten zijn gevraagd mee te doen aan een kort onderzoek wat ongeveer 20 minuten van hun tijd in beslag zou nemen. De enquetes zijn afgenomen binnen een periode van drie weken.

3.3 Respondenten

Er hebben 49 respondenten meegedaan aan het onderzoek. De leeftijden liepen uiteen van 13 tot en met 62 jaar, met een gemiddelde van 31 jaar *(SD = 14,1)*. Van de 49 respondenten waren er 16 man (33%) en 33 vrouw (67%). Er is gekozen voor een meerderheid aan vrouwelijke respondenten, omdat de gebruikte *forumthreads* van een vrouwelijk forum afkwamen. Hierdoor waren de onderwerpen ook vrouwelijker waardoor het waarschijnlijk vrouwen meer aansprak dan mannen.

3.4 Verwerking Gegevens

Voor het verwerken en analyseren van de gegevens is gebruik gemaakt van SPSS. Er is gekeken hoe de respondenten de *forumposts* classificeerden en naar de mate van overeenstemming. Dit is gemeten met een Krippendorff Alpha, die via een online applicatie is uitgevoerd: ReCal3[4]. De Krippendorff Alpha meet de betrouwbaarheid van het classificatieschema door te controleren in hoeverre mensen het met elkaar eens zijn over de klassen en dus de *forumposts* in dezelfde klassen

4 ReCal3: Reliability for 3+ Coders, *dfreelon*, geraadpleegd op 1 juni 2014, van http://dfreelon.org/utils/recalfront/recal3/#doc

plaatsen. Als verschillende mensen consistent dezelfde klassen gebruiken, kunnen we opmaken dat de mensen het classificatieschema en de verschillende klassen op eenzelfde manier begrijpen (Artstein & Poesio, 2008). De gebruikte applicatie ReCal3 houdt rekening met *multi-label* gevallen: respondenten die een enkele *forumpost* in meerdere klassen plaatsen. ReCal3 vergelijkt namelijk continu de wel en niet gekozen klassen van twee respondenten en vervolgens worden de scores van alle paren bij elkaar opgeteld. Daarentegen kan ReCal3 slechts één variabele per keer berekenen: de resultaten met betrekking tot de klassenindeling van de *forumposts* moeten dus per *forumpost* ingevoerd worden, waarna de applicatie de *agreement* (overeenstemming) per *forumpost* berekent. Zoals in Tabel 1 te zien is, geven de resultaten onder de 0 een slechte overeenkomst weer, de resultaten tussen de 0,01 en 0,20 een lichte overeenkomst, de resultaten tussen de 0,21 en 0,40 een tamelijke overeenkomst, de resultaten tussen de 0,41 en 0,60 een gematigde overeenkomst, de resultaten tussen de 0,61 en 0,80 een substantiële overeenkomst en de resultaten tussen de 0,81 en 1,00 een bijna perfecte overeenkomst. Resultaten boven de 0,4 worden over het algemeen als adequaat beschouwd (Artstein & Poesio, 2008).

Om ReCal3 te gebruiken moeten de resultaten omgeschaald worden naar een binaire schaal: een code die slechts bestaat uit twee getallen, meestal de nul (0) en de één (1) (Encyclo[5], "binaire schaal", 2014). In dit geval geven de "nullen" aan dat de respondent een bepaalde klasse niet heeft gekozen en de "enen" geven aan dat de respondent een klasse wel heeft gekozen. Door een binaire schaal te gebruiken worden ook de klassen die niet worden gekozen meegenomen in de berekening van de mate van overeenkomst; het niet kiezen van een klasse is immers ook een vorm van overeenkomst. ReCal3 vergelijkt continu de wel en niet gekozen klassen van twee respondenten: als respondent één bijvoorbeeld aan een *forumpost* klasse A en B toekent (een *multi-label* situatie) en respondent twee aan diezelfde *forumpost* slechts klasse A toekent, is er sprake van een hoge mate van overeenkomst omdat deze respondenten het enkel over klasse B niet eens zijn. Vervolgens worden de scores van alle paren bij elkaar opgeteld. De applicatie zorgt er echter wel voor dat grote nummers van afwezigheid (niet gekozen klassen) gecompenseerd worden tijdens de berekening van de Krippendorff Alpha. Dit heeft als gevolg dat het percentage van overeenkomst soms vrij hoog kan zijn, terwijl de waarden van de Krippendorff Alpha laag zijn. Hierdoor wordt de Krippendorff Alpha gebaseerd op de gekozen klassen en niet beïnvloed door het aanzienlijk hogere aantal van niet gekozen klassen.

Tabel 1

5 Binaire Schaal op Encyclo, http://www.encyclo.nl/begrip/binaire%20code

Alpha Waarden en Kracht van Overeenkomst volgens Landis en Koch (1977), uit Artstein en Poesio (2008)

k	Interpretation
< 0	Poor agreement
0.01 – 0.20	Slight agreement *
0.21 – 0.40	Fair agreement **
0.41 – 0.60	Moderate agreement ***
0.61 – 0.80	Substantial agreement ****
0.81 – 1.00	Almost perfect agreement *****

Tot slot zijn de geschreven samenvattingen handmatig en met SPSS geanalyseerd. Ten eerste is handmatig gekeken naar de manier waarop is samengevat: extractief of abstractief. Hierbij zijn alle samenvattingen met één of meer woorden die niet in de *forumposts* voorkomen beoordeeld als zijnde abstractief geschreven, omdat deze samenvattingen dus in eigen woorden zijn geschreven. Alle samenvattingen die daarentegen slechts bestaan uit een opsomming van enkele *forumposts*, zijn beoordeeld als zijnde extractief samengevat. Daarnaast is met SPSS geanalyseerd welke *forumposts* in de samenvattingen gebruikt zijn: dit is gedaan door uit elke *forumpost* alle belangrijke woorden te halen en te tellen hoe vaak deze in de samenvattingen gebruikt zijn. Wanneer er sprake was van overlap tussen de *forumposts* of er geen woorden in stonden die de *forumpost* kon onderscheiden van de anderen, is handmatig gecontroleerd of deze *forumpost* in de samenvattingen is verwerkt. Over de relevantiescores van de *forumposts* is vervolgens een t-toets tussen twee onafhankelijke metingen uitgevoerd. Hiervoor zijn per *forumpost* twee groepen gemaakt, namelijk aan de ene kant de respondenten die de *forumpost* wel in hun samenvatting hebben gebruikt en aan de andere kant de respondenten die de *forumpost* niet in hun samenvatting hebben gebruikt. Deze t-toets is uitgevoerd om te meten of er een verband bestaat tussen de relevantiescores die de respondenten de *forumposts* gaven en of ze de *forumposts* vervolgens in hun samenvattingen gebruikt hebben.

3.5 Dataset

De resultaten van de relevantiebeoordelingen, de classificeringsvragen en de gecreëerde samenvattingen zijn samengevoegd tot een dataset. Deze dataset is online te vinden op FileFactory[6], een website waar gratis bestanden gedeeld kunnen worden. Hier kan in vervolgonderzoek van gebruik gemaakt worden.

6 Filefactory, http://www.filefactory.com/file/5xcxjt0vf37j/Forum_Berichten_01.sav

Hoofdstuk 4. Resultaten

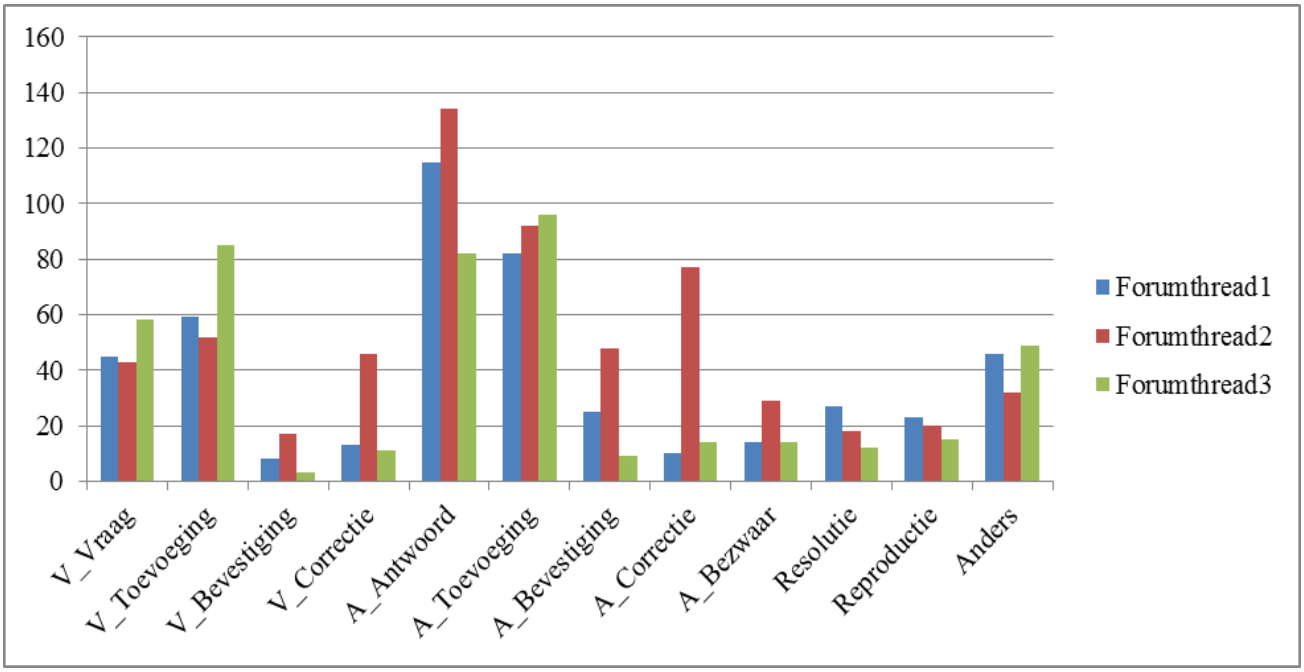
4.1 Classificatie

In deze eerste paragraaf zullen de resultaten van de classificatievragen gepresenteerd worden, omdat deze resultaten nodig zijn om antwoord te geven op de eerste deelvraag: In hoeverre komen mensen overeen in het classificeren van *forumposts*uit een *forumthread?*

In Figuur 1 en Tabel 2 zijn overzichten van de gebruikte klassen per *forumthread* te zien. Enkele klassen zijn aanzienlijk meer gebruikt dan anderen, zoals de Vraag-Toevoeging, Antwoord-Antwoord en Antwoord-Toevoeging klassen. Daarnaast zijn er ook enkele klassen juist aanzienlijk minder gebruikt dan anderen, zoals de Vraag-Bevestiging klasse die in *forumthread* 1 8 keer gebruikt is, in *forumthread* 2 17 keer gebruikt is en in de laatste *forumthread* 3 keer gebruikt is. De Antwoord-Antwoord klasse is het meest gebruikt met een totaal van 331 keer. Daartegenover is de Vraag-Bevestiging klasse het minst gebruikt met een totaal van 28 keer.

Figuur 1

Het Aantal Keer dat een Bepaalde Klasse in een Forumthread is Gekozen



Tabel 2

Het Aantal Keer dat een Bepaalde Klasse in een Forumthread is Gekozen

	Forumthread 1	Forumthread 2	Forumthread 3	Totaal
Vraag-Vraag	45	43	58	**146**
Vraag-Toevoeging	59	52	85	**196**
Vraag-Bevestiging	8	17	3	**28**
Vraag-Correctie	13	46	11	**70**
Antwoord-Antwoord	115	134	82	**331**
Antwoord-Toevoeging	82	92	96	**270**
Antwoord-Bevestiging	25	48	9	**82**
Antwoord-Correctie	10	77	14	**101**
Antwoord-Bezwaar	14	29	14	**57**
Resolutie	27	18	12	**57**
Reproductie	23	20	15	**58**
Anders	46	32	49	**127**

De Krippendorff Alpha is gebruikt om te meten in hoeverre de respondenten het met elkaar eens waren tijdens het classificeren van de *forumposts*. De resultaten zijn vervolgens beoordeeld met behulp van een interpretatie model zoals gepresenteerd in het artikel van Artstein en Poesio (2008) (Tabel 1).

In Tabel 3 en Tabel 4 staan de resultaten van de Krippendorff Alpha. Deze resultaten geven de overeenkomst van alle respondenten weer op elk individueel forumbericht. De eerste rij geeft de resultaten weer van de eerste berichten in de drie *forumthreads*, namelijk de berichten van de mensen die de *forumthreads* initieerden door een vraag te stellen. In het classificatieschema stond onder andere vermeld dat de klasse "VRAAG-VRAAG" bestemd was voor de eerste *post* in een *forumthread*. De eerste *forumposts* in de in dit onderzoek gebruikte *forumthreads* horen dus per definitie in de "VRAAG-VRAAG"-klasse (dit is overigens de enige *forumpost* waarvoor vaststaat welke klasse eraan toegekend moest worden). Deze vraag is erin gezet als een soort controlevraag: hebben de respondenten daadwerkelijk het classificatieschema goed gelezen en begrepen? In Tabel 3 is te zien dat de eerste *posts* inderdaad de hoogste mate van overeenkomst vertonen en daarnaast alle drie boven de 0,4 komen: namelijk opeenvolgend een gematigde overeenkomst (0,449), een substantiële overeenkomst (0,649) en een bijna perfecte overeenkomst (0,833). Hieruit is ook op te maken dat de respondenten het schema beter begrepen naarmate ze er meer ervaring mee hadden: de mate van overeenkomst loopt wat betreft de eerste *forumpost* op van *forumthread* 1 naar *forumthread* 3. De mate van overeenkomst in de overige *posts* varieert van een lichte overeenkomst

tot een gematigde overeenkomst. 83% van de resultaten ligt onder de 0,4, wat volgens Artstein en Poesio (2008) de grens is vanaf wanneer de overeenkomst adequaat is.

Tabel 3

Krippendorff Alpha Resultaten (Alpha per forumpost)

	Forumthread 1	**Forumthread 2**	**Forumthread 3**
Forumpost 1	*** 0,449	**** 0,649	***** 0,833
Forumpost 2	** 0,204	* 0,176	** 0,374
Forumpost 3	** 0,298	* 0,125	* 0,142
Forumpost 4	* 0,070	* 0,119	** 0,340
Forumpost 5	** 0,328	* 0,180	* 0,197
Forumpost 6	* 0,179	* 0,126	** 0,214
Forumpost 7	** 0,211	** 0,270	*** 0,516
Forumpost 8	** 0,341	* 0,155	* 0,292

Tabel 4

Krippendorff Alpha Resultaten (gemiddelde Alpha per forumthread)

	Forumthread 1	**Forumthread 2**	**Forumthread 3**
Gemiddelde Alpha	0,233	0,164	0,296

In Tabel 4 zijn de gemiddeldes van de Krippendorff Alpha per *forumthread* gegeven. Hierbij zijn de eerste *forumposts* niet meegenomen in de berekening, omdat deze *posts* de beginvraag/topicstart bevatten.

4.2 Relevantie

De resultaten in deze paragraaf hebben betrekking op de tweede deelvraag: In hoeverre kunnen reacties op vragen gesteld op internetfora door mensen worden beoordeeld als relevant of niet relevant?

In het onderzoek is de respondenten gevraagd de *forumposts* op relevantie te beoordelen met behulp van een vijfpuntsschaal. In Tabel 5 is te zien in hoeverre de respondenten de *forumposts* als zijnde relevant hebben beoordeeld. De score kon uiteenlopen van één tot vijf, waarbij een lager cijfer voor een hogere relevantie staat. Voor alle drie de *forumthreads* geldt dat de eerste twee opties, namelijk "ik vind deze post relevant" en "ik vind deze post enigszins relevant", het meeste zijn gebruikt. Bij vijf *forumposts* is een gehele beoordelingsoptie niet gebruikt en dit was bij alle vijf de *posts* de optie "ik vind deze post niet relevant".

Tabel 5

De Gemiddelde Relevantiescores van de Individuele Forumposts voor de Forumthreads (met de standaarddeviatie tussen haakjes)

	Forumthread 1	**Forumthread 2**	**Forumthread 3**
Post 2	1,96 (1,23)	3,29 (1,61)	2,84 (1,34)
Post 3	2,73 (1,32)	2,35 (1,11)	1,92 (1,13)
Post 4	2,12 (1,18)	2,59 (1,08)	3,47 (1,45)
Post 5	1,55 (0,91)	1,41 (0,71)	2,76 (1,35)
Post 6	1,65 (0,97)	3,02 (1,30)	3,96 (1,22)
Post 7	1,96 (1,22)	1,33 (0,63)	1,35 (0,63)
Post 8	4,61 (0,91)	3,29 (1,61)	3,49 (1,39)

4.3 Samenvattingen

De resultaten in deze laatste paragraaf hebben betrekking op de derde deelvraag: In hoeverre komen door mensen gemaakte samenvattingen aan de hand van relevante reacties op vragen op internetfora overeen? Om de tweede deelvraag te kunnen beantwoorden moet geanalyseerd worden wat de respondenten hetzelfde of juist anders hebben gedaan tijdens het maken van hun samenvattingen. In deze paragraaf zal besproken worden welke *forumposts* gebruikt zijn in de samenvattingen en op welke manier de respondenten de samenvattingen hebben geschreven: abstractief of extractief.

De respondenten hebben niet alle *forumposts* gebruikt in hun samenvattingen. In Tabel 6 is te zien welke *forumposts* gebruikt zijn en in welke mate. Er is te zien dat bepaalde *forumposts* aanzienlijk meer zijn gebruikt dan anderen en enkele helemaal niet. In *forumthread* 1 zijn *forumpost* 5 (73%), 6 (92%) en 7 (67%) het meest gebruikt in de samenvattingen. *Forumpost* 3 kwam in geen enkele samenvatting voor. In de samenvattingen van *forumthread* 2 zijn *forumpost* 3 (71%) en 5 (45%) het meest gebruikt. Geen enkele *forumpost* is niet gebruikt; *forumpost* 6 kwam in twee samenvattingen voor en is hiermee het minste in de samenvattingen verwerkt (4%). Tot slot is in *forumthread* 3 *forumpost* 7 het meest gebruikt in de samenvattingen, namelijk 23 keer (47%). In deze *thread* zijn 3 *forumposts* niet door de respondenten gebruikt in de samenvattingen: *forumpost* 5, 6 en 8.

Tabel 6

Het Aantal Keer dat de Forumposts in de Samenvattingen is gebruikt, per Forumpost (in aantallen en percentages)

	Forumthread 1	**Forumthread 2**	**Forumthread 3**
Post 2	9 (18%)	6 (12%)	10 (20%)
Post 3	0 (0%)	35 (71%)	8 (16%)
Post 4	1 (2%)	8 (16%)	8 (16%)
Post 5	36 (73%)	22 (45%)	0 (0%)
Post 6	45 (92%)	2 (4%)	0 (0%)
Post 7	33 (67%)	14 (29%)	23 (47%)
Post 8	1 (2%)	5 (10%)	0 (0%)

In Kader 1 is *forumpost* 5 uit *forumthread* 1 te zien; de reactie die het meest in de samenvattingen is gebruikt, namelijk 45 keer (92%). In Kader 2 is *forumpost* 8 uit *forumthread* 3 te zien; één van de vijf reacties die geen enkele keer in de samenvattingen is voorgekomen.

Kader 1

Forumpost 5 uit Forumthread 1 – Meest Gebruikt in de Samenvattingen (92%)

"Dan zou ik binnenkort nog eens bij je ouders op de Weegschaal gaan staan. Ik denk dat deze van je vriendin een afwijking heeft ten opzichte van de Weegschaal van je ouders.
Geen zorgen maken! Ik hield toch best wat vocht vast tijdens de zwangerschap, dat was verdeeld over mijn hele lijf. Nadat ik was bevallen moest ik een dag of twee echt heel vaak plassen en was ik zo weer een paar kilo kwijt aan vocht.
Je wordt altijd wat dikker als je zwanger bent. Dat is normaal en hoef je je echt niet druk om te maken! Komt goed."

-pop1980

Kader 2

Forumpost 8 uit Forumthread 3 – In Geen Enkele Samenvatting Gebruikt (0%)

"Ik kan het niet uitstaan dat ik niet weet hoe die "rail" heet en er dus geen voorbeeld van vinden. Omaatjes hebben er altijd van die korte kanten gordijntjes aan hangen. Ze zijn flexibel, een beetje rekbaar en wit. Aan het einde zit dan zo'n dopje die je door een haakje doet waardoor hij blijft hangen."
-876756756

Over de relevantiescores van de *forumposts* is een t-toets tussen twee onafhankelijke metingen uitgevoerd, namelijk tussen de respondenten die de *forumpost* wel in hun samenvatting hebben gebruikt en de respondenten die de *forumpost* niet in hun samenvatting hebben gebruikt. Deze t-toets is uitgevoerd om te meten of er een verband bestaat tussen de relevantiescore die de respondenten de *forumposts* gaven en of ze de *forumposts* vervolgens in hun samenvattingen gebruikt hebben. Hiermee is dus gemeten of respondenten tijdens het samenvatten rekening houden met wat volgens hen de relevante *forumposts* zijn.

In Tabel 7 en 8 zijn opeenvolgend de gemiddeldes per *forumpost* en per groep te zien en vervolgens de gemiddeldes per *forumthread*. Over de *forumposts* die geen enkele keer in de samenvattingen gebruikt zijn kon geen t-toets worden uitgevoerd (*forumpost* 3 uit *forumthread* 1 en *forumposts* 5, 6 en 8 uit *forumthread* 3).

Tabel 7

Gemiddelde Relevantiescores (Significante Verschillen Benadrukt)

	Forumthread 1		**Forumthread 2**		**Forumthread 3**	
	Gebruikt	Niet gebruikt	Gebruikt	Niet gebruikt	Gebruikt	Niet gebruikt
Post 2	2,44	1,85	2,33	2,02	2,10	3,03
Post 3	-	2,73	2,31	2,43	1,50	2,00
Post 4	1,00	2,15	2,13	2,68	4,00	3,42
Post 5	**1,39**	**2,00**	**1,18**	**1,59**	-	2,73
Post 6	1,67	1,50	2,00	3,06	-	3,96
Post 7	1,94	2,00	1,07	1,43	1,17	1,50
Post 8	5,00	4,60	2,00	3,43	-	3,49

Tabel 8

Gemiddelde Relevantiescores per Forumthread

	Forumthread 1		**Forumthread 2**		**Forumthread 3**	
	Gebruikt	Niet gebruikt	Gebruikt	Niet gebruikt	Gebruikt	Niet gebruikt
Totaal	2,24	2,40	1,86	2,38	2,19	2,88

Twee t-toetsen gaven een significant resultaat, namelijk die over *forumpost* 5 uit *forumthread* 1 en *forumpost* 5 uit *forumthread* 2.

De Levene's Test is eerst uitgevoerd om te kijken of de varianties van de groepen gelijk zijn. Voor *forumpost* 5 uit *forumthread* 1 mogen de varianties van beide groepen gelijk worden geacht (p=0,959). De relevantiescores van de respondenten over *forumpost* 5 uit *forumthread* 1 die de *forumpost* gebruikt hebben in de samenvattingen (M=1,39, SD=0,84) verschillen significant van de relevantiescores van de respondenten die *forumpost* 5 niet gebruikt hebben in hun samenvattingen (M=2,00, SD=1,00), t=2,14; p=0,037, 95%*CI* [0,04, 1,19].

Voor *forumpost* 5 uit *forumthread* 2 mogen de varianties van beide groepen niet als gelijk worden geacht (p=0,005). Bij deze resultaten is dus gekeken naar de *equal variances not assumed* cijfers. De relevantiescores van de respondenten over *forumpost* 5 uit *forumthread* 1 die de *forumpost* gebruikt hebben in hun samenvattingen (M=1,18, SD=0,50) verschillen significant van de relevantiescores van de respondenten die *forumpost* 5 niet gebruikt hebben in hun samenvattingen (M=1,59, SD=0,80), t=2,20; p=0,033, 95%*CI*[0,03, 0,79]. De overige t-toetsen hebben geen significant verschil aangetoond.

Naast het gebruik van de *forumposts* in de samenvattingen, is ook geanalyseerd hoe de respondenten de samenvattingen hebben geschreven: ofwel extractief ofwel abstractief. In Tabel 9 staat weergeven hoeveel procent van de respondenten de *forumthreads* extractief en hoeveel procent de *forumthreads* abstractief hebben samengevat. In alle drie de *forumthreads* heeft meer dan 90% van de respondenten de *forumthread* abstractief samengevat, wat inhoudt dat de samenvatting in ieder geval enigszins in eigen woorden is gezet. In totaal hebben 9 van de 49 mensen in ieder geval één keer extractief samengevat, wat inhoudt dat de samenvatting niet in eigen woorden is gezet, waarvan 3 mensen dit meerdere keren hebben gedaan. 33% van de respondenten die extractief heeft samengevat, heeft deze techniek dus meerdere keren toegepast.

Tabel 9

Aantal Samenvattingen Extractief en Abstractief Geschreven (in percentages)

	Extractief	**Abstractief**
Forumthread 1	6%	94%
Forumthread 2	10%	90%
Forumthread 3	10%	90%
Totaal	9%	91%

De respondenten gebruikten verschillende technieken tijdens het samenvatten van de *forumthreads*. Over het algemeen werden de berichten die werden gezien als zijnde van belang opgesomd en tot een lopend verhaal gevormd. Er waren echter ook respondenten die puntsgewijs hebben samengevat, zoals te zien is in Kader 3 en Kader 4. Er waren ook enkele respondenten die de *forumthreads* niet hebben samengevat, maar een oplossing op de vraagstelling gaven of een reactie op de discussie in de *forumthreads*. Een voorbeeld van een abstractief geschreven samenvatting is te zien in Kader 5. Hier is in te zien dat de respondent eigen woorden heeft toegevoegd die niet in de *forumthread* stonden, bijvoorbeeld "de forumthread", "betreft" en "verschillende mensen".

Kader 3

Samenvatting van Forumthread 3 door Respondent 20 – Extractief Geschreven

"Vraag: Man/vrouw heeft problemen met de gleuf in de deur die door moet gaan als brievenbus. Persoon wil graag post opgevangen hebben voordat het de grond raakt. Heeft de oplossing al gezien, maar vindt dit te duur. Hoe kan iets goedkoop gemaakt worden?"

Kader 4

Samenvatting van Forumthread 3 door Respondent 17 – Extractief Geschreven

"V: Kan ik goedkoop een postzak voor bij de brievenbus maken? A: Misschien een linnen zak gebruiken en die vastmaken aan een rail dmv haakjes."

Kader 5

Samenvatting van Forumthread 1 door Respondent 10 – Abstractief Geschreven

"De forumthread betreft het aankomen tijdens een zwangerschap en wat hierbij normaal is. Verschillende mensen geven aan dat het verstandig is om op een goede, en altijd dezelfde weegschaal te gaan staan, en dat het heel normaal is om tijdens je zwangerschap aan te komen."

Hoofdstuk 5. Discussie

5.1 In hoeverre komen mensen overeen in het classificeren van *forumposts* uit een *forumthread?*

Ten eerste is onderzocht hoe mensen *forumposts* classificeren en in hoeverre deze keuze overeenkomt. Zoals gebleken uit de resultaten was er geen hoge mate van overeenkomst: 83% van de resultaten lag onder de grens vanaf wanneer de mate van overeenkomst adequaat te noemen is. Daarentegen is de mate van overeenkomst bij de eerste *posts* van de *forumthreads* bij alle drie adequaat en bij de laatste zelfs "bijna perfect". Hieruit valt te concluderen dat de respondenten het classificatieschema wel degelijk goed lazen, alleen de klassen anders begrepen. In dit onderzoek is een aanzienlijk grotere groep respondenten gebruikt dan in voorgaande onderzoeken met een soortgelijk classificatieschema. De verklaring voor de lage mate van overeenkomst zou dus kunnen liggen bij het classificatieschema zelf, wat aangeeft dat een dergelijk classificatieschema niet goed werkt bij grote groepen. Immers, hoe groter de groep, hoe groter de afwijkingen tussen de classificatiekeuzes zullen zijn. Daarnaast zou het ook kunnen zijn dat het classificeren van *forumposts* uit een *forumthread* voor mensen een moeilijke taak is, waardoor het toekennen van een bepaalde klasse aan een *post* geen logische handeling is die iedereen hetzelfde uitvoert.

Tijdens het onderzoek is overwogen om een extra analyse uit te voeren om de mate van overeenkomst te berekenen, bijvoorbeeld door de respondenten die meerdere klassen toekenden aan één *forumpost* als een aparte klasse te beschouwen. Er is echt voor gekozen dit niet te doen, omdat de resultaten van de Krippendorff Alpha berekend door de eerder genoemde gebruikte applicatie een goede afspiegeling geven van de mate van overeenkomst tussen de mensen. Ook het gebruik van de binaire schaal, waarbij ook de niet gekozen klassen mee worden genomen in de berekening (het niet kiezen van een klasse is immers ook een vorm van overeenkomst), had geen invloed op de uiteindelijke waarden van de Krippendorff Alpha. De applicatie compenseerde de hoge waarden van de afwezige klassen.

5.2 In hoeverre kunnen reacties op vragen gesteld op internetfora door mensen worden beoordeeld als relevant of niet relevant?

Tevens is onderzocht hoe relevant de mensen de *forumposts* voor de *forumthread* vonden. De meeste *forumposts* werden beoordeeld met "relevant" en "enigszins relevant". De beoordelingsoptie "niet relevant" werd in vijf *forumposts* zelfs door geen enkele respondent gebruikt. Hieruit valt te concluderen dat mensen een reactie in een *forumthread* niet snel als irrelevant zien; de meeste mensen vinden dat het toch altijd iets toevoegt aan de context. De reacties op vragen gesteld op internetfora kunnen dus over het algemeen beoordeeld worden als relevant voor het onderwerp.

5.3 In hoeverre komen door mensen gemaakte samenvattingen aan de hand van relevante reacties op vragen op internetfora overeen?

Om een antwoord op deze subvraag te vinden zijn de samenvattingen met elkaar vergeleken. Ten eerste is gekeken naar de gebruikte *forumposts* in de samenvattingen. In de resultaten is te zien dat bepaalde *forumposts* aanzienlijk meer zijn gebruikt dan anderen en enkele helemaal niet. Dit geeft aan dat mensen een gemeenschappelijke voorkeur hebben voor bepaalde reacties. *Forumpost* 5 uit *forumthread* 1 is het meest gebruikt in de samenvattingen. Behalve een antwoord op de vraag, staat er in deze reactie ook informatie over de eigen ervaring van de schrijver met betrekking tot het onderwerp. Daarnaast wordt er ook nog bemoedigend gesproken en de gehele reactie is een stuk positiever geschreven dan de overige reacties. Mensen vinden het dus fijn als er een uitgebreide reactie wordt gegeven. Toch hoeft een uitgebreide reactie niet gelijk goed te zijn, zoals te zien is in *forumpost* 8 uit *forumthread* 3, die juist in geen enkele samenvatting is gebruikt. Hoewel deze reactie vrij uitgebreid is, staat er geen antwoord in op de vraag en er wordt ook niet gesproken over de eigen ervaring omtrent het onderwerp. Deze *forumpost* is daarnaast door 26 mensen beoordeeld als zijnde irrelevant en slechts door 12 als relevant (11 mensen vonden de reactie neutraal). Daarentegen is de eerder genoemde *forumpost* 5 uit *forumthread* 1 door 43 mensen beoordeeld als zijnde relevant en slechts door 4 mensen als enigszins irrelevant (2 mensen vonden de reactie neutraal). Deze resultaten komen bijna compleet overeen met de hoeveelheid mensen die de reactie ook daadwerkelijk in hun samenvatting hebben gebruikt: 43 mensen vonden de reactie relevant en 45 mensen hebben de reactie in hun samenvatting toegevoegd.

Er is ook onderzocht hoe de mensen de samenvattingen schreven: abstractief of extractief. In de resultaten is te zien dat de grote meerderheid van de mensen *forumthreads* abstractief samenvat. Hierbij worden de samenvattingen in eigen woorden gezet, in plaats van slechts woorden te gebruiken die in de tekst voor komen. De mensen die extractief samenvatten, gebruiken tijdens het schrijven van de samenvatting enkel deze strategie, terwijl een deel van de mensen die abstractief samenvatten, ook deels extractief samenvatten: in dit geval wordt informatie uit de brontekst alsnog opgesomd, alleen worden hier en daar eigen woorden toegevoegd om er een lopend verhaal van te maken. Hier moet rekening mee worden gehouden tijdens het ontwerpen van een automatisch samenvattingssysteem: mensen schrijven zelf vaak abstractieve samenvattingen en zijn deze dus meer gewend om te lezen. Wanneer er echter voor gekozen wordt om het samenvattingssysteem tevens abstractief te laten samenvatten, wordt het trainen van het systeem een moeilijkere taak. Eerder is namelijk al beschreven hoe het momenteel voor samenvattingssystemen nog moeilijk (maar niet onmogelijk) is om abstractief samen te vatten. 33% van de mensen die de *forumthreads* wel extractief heeft samengevat, heeft dit ook bij meerdere *forumthreads* toegepast. Hoewel extractief samenvatten dus bijna niet wordt gebruikt, is het wel een techniek die mensen als ze deze

gebruiken, mogelijk ook vaker toepassen. Daarnaast vatten de meeste mensen de fora samen door de berichten die zij als belangrijk achtten in de context op te sommen en daar een lopend verhaal van te maken. Slechts enkele mensen varieerden door bijvoorbeeld een puntsgewijze samenvatting te maken. Tot slot waren er ook mensen die geen daadwerkelijke samenvatting schreven, maar een oplossing op de vraagstelling of een mening gaven over de discussie in de *forumthreads*. Dit kan zijn omdat ze de vraag verkeerd begrepen of te snel gelezen hebben. Aangezien de vraag door het grootste deel echter wel goed werd ingevuld, lijkt het erop dat de mensen op dit punt in de vragenlijst geen zin meer hadden om zelf nog een tekst te typen.

5.4 Hoe maken mensen een samenvatting van een online discussie terwijl ze de informatie uit de discussie classificeren en het relevante selecteren?

Met de resultaten van de drie subvragen kan nu een antwoord gevormd worden op de hoofdvraag van dit onderzoek: hoe maken mensen een samenvatting van een online discussie terwijl ze de informatie uit de discussie classificeren en het relevante selecteren? Ten eerste voeren mensen de classificatietaak in grote mate anders uit dan elkaar. Mensen begrijpen het classificatieschema maar interpreteren het op een andere manier en classificeren vervolgens verschillend van elkaar. Daarnaast gebruiken de meeste mensen de reacties die ze het meest relevant voor het onderwerp vinden in een samenvatting. Over het algemeen hebben mensen een gemeenschappelijke voorkeur voor reacties en zijn ze het er over eens welke reacties meer of minder relevant zijn voor het onderwerp, waardoor ze ook een gemeenschappelijke voorkeur hebben voor de reacties die ze in een samenvatting willen gebruiken. Mensen schrijven hun samenvatting dus grotendeels op basis van hun eigen beoordeling van de reacties in de online discussie, in plaats van dat ze meer of minder gebruik maken van bepaalde klassen. Het is dus handig een classificatieschema te gebruiken tijdens een online discussie, omdat dit de reacties op een logische wijze organiseert, maar dit lijkt geen noodzaak te zijn voor de samenvattingstaak. Tot slot vatten de meeste mensen abstractief samen. Hoewel dit niet betekent dat mensen ook liever een abstractieve samenvatting lezen dan een extractieve samenvatting, zijn ze dit wel meer gewend.

Aanbevelingen

In dit onderzoek is een dataset opgebouwd bestaande uit handmatig geschreven samenvattingen en gegevens omtrent relevantie- en classificeervragen. Deze dataset kan in vervolgonderzoek gebruikt worden.

Het zou interessant zijn om in de toekomst een vervolgonderzoek op deze studie te doen en te onderzoeken of met behulp van de verzamelde resultaten in de dataset omtrent het handmatig samenvatten een automatisch samenvattingssysteem te trainen is. Daarbij zou er eventueel voor gekozen kunnen worden de samenvattingen niet abstractief maar extractief te genereren, omdat dit momenteel realistischer is.

Uit de resultaten van het onderzoek is gebleken dat het classificatieschema niet goed werkt bij grote groepen respondenten: de mate van overeenkomst was immers vrij laag. In een vervolgonderzoek zou eventueel een nieuw classificatieschema kunnen worden opgebouwd waarbij gebruik wordt gemaakt van de dataset, specifiek voor grotere groepen respondenten, of er zou onderzocht kunnen worden of classificatieschema's überhaupt op grotere groepen mensen zijn toe te passen.

Referenties

Alguliev, R., & Aliguliyev, R. (2009). Evolutionary Algorithm for Extractive Text Summarization. *Intelligent Information Management, 1*(2), p 128-138.

Aone, C., Okurowski, M. E., Gorlinksy, J., & Larsen, B. (1999). A Trainable Summarizer with Knowledge Acquired from Robust NLP Techniques. *Advances in automatic text summarization*, p 71.

Artstein, R., & Poesio, M. (2008). Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics, 34*(4), p 555-596.

Barzilay, R., & Elhadad, M. (1997). Using lexical chains for text summarization. *Proceedings of the ACL workshop on intelligent scalable text summarization, 17*(1), p 10-17).

Brown, A.L. (1983). Macrorules for Summarizing Texts: The Development of Expertise. *Journal of Verbal Learning and Verbal Behavior, 22*, p 1-14.

Brown, A. L., Day, J. D., & Jones, R. S. (1983). The development of plans for summarizing texts. *Child Development, 54*(4), p 968-979.

Carenini, G., & Cheung, J. C. K. (2008). Extractive vs. NLG-based abstractive summarization of evaluative text: the effect of corpus controversiality. *Proceedings of the Fifth International Natural Language Generation Conference*. Association for Computational Linguistics, Stroudsburg, PA, USA, p 33-41.

Copeck, T., & Szpakowicz, S. (2004). Vocabulary agreement among model summaries and source documents. *Les actes de Document Understanding Conference (DUC)*.

Cormack, R.M. (1971). A Review of Classification. *Journal of the Royal Statistical Society, 134*(3), p 321-367.

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on, 13*(1), 21-27.

Eppler, M. J., & Mengis, J. (2004). The concept of information overload: A review of literature from organization science, accounting, marketing, MIS, and related disciplines. *The information society, 20*(5), p 325-344.

Goldstein, J., Kantrowitz, M., Mittal, V., & Carbonell, J. (1999). Summarizing Text Documents: Sentence Selection and Evaluation Metrics. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '99). ACM, New York, NY, USA, p 121-128.

Gupta, V., & Lehal, G. S. (2010). A Survey of Text Summarization Extractive Techniques. *Journal of Emerging Technologies in Web Intelligence, 2*(3), p 258-268.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM,

New York, NY, USA, p 168-177

Jackson, P., & Moulinier, I. (2007). Natural language processing for online applications: Tekst retrieval, extraction and categorization (Vol. 5). John Benjamins Publishing.

Jones, K.S. (1999). Automatic summarizing: factors and directions. In I. Mani & M. Maybury (Ed.), *Advances in automatic text summarisation* (p 1-12). Cambridge MA: MIT Press.

Kim, S. N., Wang, L., & Baldwin, T. (2010). Tagging and linking web forum posts. *Proceedings of the Fourteenth Conference on Computational Natural Language Learning*, p. 192-202. Association for Computational Linguistics.

Knight, K., & Marcu, D. (2002). Summarization beyond sentence extraction: a probabilistic approach to sentence compression. *Artificial Intelligence, 139* (1), p 91-107.

McKeown, K., Klavans, J., Hatzivassiloglou, V., Barzilay, R., & Eskin, E. (1999). Towards multidocument summarization by reformulation: Progress and prospects. In: Proceedings of AAAI-99, Orlando, FL, p 453–460.

McKeown, K., Passonneau, R. J., Elson, D. K., Nenkova, A., & Hirschberg, J. (2005). Do summaries help? A task-based evaluation of multi-document summarization. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, p 210-217.

Nenkova, A., & Vanderwende, L. (2005). The impact of frequency on summarization. *Microsoft Research, Redmond, Washington, Tech. Rep. MSR-TR-2005-101*.

Schiffman, B., Nenkova, A., & McKeown, K. (2002). Experiments in multidocument summarization. *Proceedings of the second international conference on Human Language Technology Research*, Morgan Kaufmann Publishers Inc, p 52-58.

Teufel, S., & Moens, M. (1997). Sentence extraction as a classification task. *Proceedings of the ACL, 97*, p 58-65.

Weinberger, K., Blitzer, J., & Saul, L. (2006). Distance metric learning for large margin nearest neighbor classification. *Advances in neural information processing systems, 18*.

Bijlage 1

Forumberichten

Forumthread één

10 kg aangekomen in 2 maanden tijd

Ok, uhmm... HELP dus! Kan dit?

Ik weeg normaal +/- 58 kg, kan al eens een kilootje meer of minder zijn, niets speciaals.

Ik ben nu 8,5 weken zwanger en ging daarstraks op de weegschaal staan bij een vriendin thuis: 70 kg (met kleren en schoenen, dus ik trek er ruim genomen nog 2 kg af). WTF?! Dat is toch absurd veel?

Ik heb wel al een klein buikje, maar al mijn kleren passen nog, enkel de kleinste broeken die ik heb zitten wat strakker. Ik ben er echt niet goed van.

Het was trouwens wel een flutweegschaaltje, zo'n paarse met als opdruk 'Deze weegschaal is niet verantwoordelijk voor het hoge cijfer' ofzoiets. Bij mijn zoontje geeft hij echter 16 kg aan. Zal niet helemaal precies zijn, maar ook niet veel ernaast.

Kan dit? 10 kg erbij in 2 maanden tijd terwijl al mijn kleren nog passen, ik vind het gek, maar het heeft me heel hard laten schrikken.

-inimommy

Ik zou eerst op een goede weegschaal gaan staan voordat je gaat panikeren

-bootje_op_de_golven

"Hoe vaak weeg je je? Je hebt zelf geen Weegschaal, hoe kan je dan weten hoeveel je weegt?"

-pop1980

"Pop, ik weeg me zo nu en dan bij mijn ouders thuis. Het zit al jaren op 58 kilo, dus ik kan het ook perfect aanvoelen aan mijn kleren. Je ziet het ook echt niet aan me, behalve dat ik een buikje heb. Daarom vind ik het ook zo raar."

-inimommy

"Weeg altijd op dezelfde weegschaal, dan weet je precies hoeveel er bij komt. En als het dan nog steeds 'veel' is, niet stressen."

-lemoos2

"Dan zou ik binnenkort nog eens bij je ouders op de Weegschaal gaan staan. Ik denk dat deze van je

vriendin een afwijking heeft ten opzichte van de Weegschaal van je ouders.
Geen zorgen maken! Ik hield toch best wat vocht vast tijdens de zwangerschap, dat was verdeeld over mijn hele lijf. Nadat ik was bevallen moest ik een dag of twee echt heel vaak plassen en was ik zo weer een paar kilo kwijt aan vocht.
Je wordt altijd wat dikker als je zwanger bent. Dat is normaal en hoef je je echt niet druk om te maken! Komt goed."

-pop1980

"Gewoon nogmaals bij je ouders wegen.
En misschien houd je extra vocht vast. Dat verdwijnt vanzelf na de bevalling."
-Java

"Fopweegschaal.."
-masque

Forumthread twee
Doorslikken?
Ik ben een beetje op dieet. Dat wil zeggen zo min mogelijk snoepen en alcohol.
Nu is 1 van mijn beste vriendinnen jarig en ik weet dat zij altijd vreselijk veel in huis haalt qua happen.
Ik ga hier weerstand tegen proberen te bieden,en bedacht mij om alles wel te eten maar niet door te slikken. Zij heeft al kersenbonbons in huis gehaald,waar ik normaliter bergen van op kan.
Wanneer ik er wel op sabbel maar niet doorslik krijg ik toch veel minder calo s binnen. En neem ik toch iets van haar traktaties.
Of is dit heel raar bedacht?
-kerriebredsjo

"Zoals je zelf al zegt.. Dat is heel raar bedacht!
Neem dan gewoon van alles 1 en laat het daarbij. Wat wou je met de afgelebberde stukken doen??"
-whatcomesaround

"Ja, dit is volslagen belachelijk. Hoe zie je dat voor je? Te midden van de visite uitgekauwde bonbons uitspugen? Hou jezelf gewoon in en prop je niet helemaal vol."
-LillyFeeee

"In mn tassie. Zakdoek voor de mond en in mn tas laten glijden.
Ik weet zeker als ik zeg dat ik op dieet ben,dat ze beginnen te blaxen dat het niet nodig is. En ik vind van wel en wil de aandacht er ook niet op vestigen. Hmmm..."
-kerriebredsjo

"Ik denk juist dat als je op dieet bent, je jezelf af en toe wel wat lekkers moet toestaan. Zo houd je het naar mijn idee ook langer vol. Een verjaardag is daar een goed voorbeeld van. Wel wat nemen, maar gewoon niet zoveel. Je hoeft je niet te verdedigen en je komt er echt niet ineens kilo's van aan."
-Joss28

"Dat doen ze ook bij proeverijen zo anders werden die mensen vanwege hun beroep baggervet, zat van de koffie of lamlazarus van de wijn.
Maar een heel pak koeken of bonbons zo opeten is een heel gedoe waar je gauw zat van krijgt en dan berg je het toch maar de kast in na een of twee koeken."
-blommit

"Je hoeft je toch niet te verdedigen. Gewoon niet over beginnen en als ze erom vraagt waarom je geen bonbons neemt zeg je kort dat je wil afvallen. Geen discussie aangaan. Als ze zegt dat is toch niet nodig? Gewoon antwoorden: dank je voor het compliment.
Geen discussie aangaan, hoef je je ook niet te verdedigen."
-floriana

"Ik zou dat in ieder geval niet in het openbaar doen"
-blommit

Forumbericht drie

Inspiratie nodig bij maken verticale brievenbuszak

Ik wil dus een brievenbus zak. Ik heb een verticale brievenbusgleuf aan de zijkant van mijn voordeur op, ongeveer 160 hoog. Deur gaat naar binnen open (brievenbus zit aan de kant van de scharnieren) dus een bak is geen optie. Nu zag ik ze op internet maar ik vind ze vrij prijzig. Heeft iemand ideeën over hoe ik het zelf kan maken voor een fractie van de prijs?
-regan

"Kun je er geen mandje onder hangen?"
-blijfgewoonbianca

"Nee, mandje gaat niet werken omdat het oppervlak te smal is (gleuf is niet voor niets verticaal) en ook kan dan de deur niet meer open. Die gaat naar binnen open en de gleuf zit aan de scharnier kant"
-regan

"Kan me er niets bij voorstellen als de bus verticaal hangt. Waarom mag de post niet op de grond vallen?"
-Aardbeienthee

"Sommige mensen kunnen/ willen niet bukken, willen niet over de post uitglijden als ze binnen komen, of de hond vreet het op."
-blijfgewoonbianca

"Of ze hebben een peuter met een geheime voorraad pennen die graag 'brieven schrijft'.
(Vorige week gebeurd hier.)."
-wendy7474

"Kan je niet de hengsels van een linnen tas knippen en deze dan met haakjes bevestigen aan je brievenbus? Of er een zoom in maken en hier dan van die ijzerdraad-met-plastic gordijn "rail" doorheen doen en dan aan de uiteinden van je brievenbus de haakjes om het vast te houden maken?"
-876756756

"Ik kan het niet uitstaan dat ik niet weet hoe die "rail" heet en er dus geen voorbeeld van vinden. Omaatjes hebben er altijd van die korte kanten gordijntjes aan hangen. Ze zijn flexibel, een beetje rekbaar en wit. Aan het einde zit dan zo'n dopje die je door een haakje doet waardoor hij blijft hangen."
-876756756

Bijlage 2

Classificatieschema

Label set

RESOURCE: Linking and dialogue act labelling of posts in web user forum data

Annotation of posts and the relation between two posts.

2 super klassen

QUESTION

ANSWER

3 singleton classes

RESOLUTION

REPRODUCTION

OTHER

QUESTION contains 4 sub-classes:

QUESTION, ADD, CONFIRMATION, CORRECTION

ANSWER contains 5 sub-classes:

ANSWER, ADD, CONFIRMATION, CORRECTION, and OBJECTION

QUESTION-QUESTION : the post contains a new question. This label is reserved for the first post in a given thread.

QUESTION-ADD : the post supplements a question by providing additional information, or asking a follow-up question.

QUESTION-CONFIRMATION : the post points out error(s) in a question without correcting them, or confirms details of the question.

QUESTION-CORRECTION: the post corrects error(s) in a question.

ANSWER-ANSWER: the post proposes an answer to a question.

ANSWER-ADD: the post supplements an answer by providing additional information.

ANSWER-CONFIRMATION: the post points out error(s) in an answer without correcting them, or confirms details of the answer.

ANSWER-CORRECTION: the post corrects error(s) in an answer.
ANSWER-OBJECTION: the post objects to an answer on experiential or theoretical grounds (e.g. It wonâ€™t work.).

RESOLUTION: the post confirms that an answer works, on the basis of implementing it.
REPRODUCTION: the post either: (1) confirms that the same problem is being experienced (by a non-initiator, e.g. Iâ€™m seeing the same thing.); or (2) confirms that the answer should work.
OTHER: the post does not belong to any of the above classes.

Annotation rules:
- a post can only be linked to a previous post
- a post can have multiple links to previous post(s)
- each first post is annotated as Question-question (no link)