Technisch rapport
SumNL corpus

Iris Hendrickx

CNTS - Language Technology Group,
University of Antwerp, Universiteitsplein 1, Antwerp
Belgium
iris.hendrickx@ua.ac.be

1 Introductie

Het SumNL corpus is ontwikkeld voor het evalueren van automatische samenvatters, met name voor vraaggestuurde multi-document summarization. Het corpus bestaat uit 30 clusters. Ieder cluster bestaat uit een onderwerp en 5-25 teksten (krantenartikelen) die relevant zijn voor het onderwerp. Voor ieder cluster zijn door 5 verschillende annotatoren samenvattingen gemaakt van verschillende grootte. Daarnaast hebben de annotatoren ook de zinnen van iedere tekst een score gegeven die hun belangrijkheid aangeeft. In de volgende sectie gaan we in op de achtergrond en het doel van het corpus. In sectie 3 geven we een gedetailleerde omschrijving van het corpus.

2 Achtergrond

Automatisch samenvatten heeft als doel een kortere versie van een tekst of van een groep gerelateerde teksten te maken waarin de meest belangrijke informatie uit de originele tekst(en) behouden blijft[2]. Het creëren van een samenvatting voor een groep teksten over hetzelfde onderwerp wordt multi-document summarization genoemd. De toepassingen van automatisch samenvatten zijn legio. Samenvattingen laten schrijven door mensen is kostbaar in tijd en inspanning. Door de digitalisering van informatie zijn steeds meer informatiebronnen beschikbaar in elektronische vorm, en automatisch gegenereerde samenvattingen kunnen helpen met het doorzoeken en aanbieden van grote hoeveelheden informatie. Multi-document summarization kan zeer nuttig zijn om redundante informatie te filteren, of om een overzicht over een bepaald onderwerp te krijgen. Bijvoorbeeld een analist die snel een overzicht wil van alles wat er vandaag is geschreven over het Amerikaanse begrotingstekort, of een journalist die een overzicht wil van de huidige theorieën over het broeikaseffect.

Om een goede samenvatting te kunnen maken van een tekst, zou een automatisch systeem in het ideale geval een volledig begrip van de tekst moeten hebben en daar een globale representatie van construeren. Dit is de manier waarop mensen een samenvatting maken en deze methode wordt aangeduid met de term *abstractie*.

Een standaard aanpak voor automatisch samenvatten is als volgt; bereken voor iedere zin in een tekst zijn belangrijkheid, sorteer de zinnen en gebruik de belangrijkste zinnen als samenvatting. Deze methode wordt *extractie* genoemd. Het cruciale punt is natuurlijk het meten van welke zinnen belangrijk zijn. De huidige systemen gebruiken vaak een statistische aanpak waarbij kenmerken zoals de positie van een zin binnen de tekst, aanwezigheid van bepaalde woorden, enz. worden gebruikt. Een systeem kan vervolgens leren welke van deze kenmerken belangrijk zijn aan de hand van een corpus van teksten en voorbeeld (door mensen gemaakte) extracts.

Het maken van samenvattingen door zinsextractie heeft echter een aantal nadelen. De samenvatting kan onsamenhangend worden doordat retorische relaties of logische structuur worden verbroken (bijvoorbeeld "aan de ene kant..." zonder "aan de andere kant..." is raar). Anaforen kunnen een probleem vormen voor de leesbaarheid. Bijvoorbeeld een persoonlijk voornaamwoord zit in een van de geselecteerde zinnen, maar de naam waar het voornaamwoord naar verwijst niet meer. Zinnen worden in hun geheel geselecteerd waardoor ook minder relevante stukken binnen een zin opgenomen in een samenvatting. En in het geval van multi-document summarization geldt het bovenstaande in nog sterkere mate.

In de laatste jaren is veel aandacht besteed aan methoden voor evaluatie van automatische samenvattingen. Tijdens de ontwikkeling van een systeem moet regelmatig getest worden of veranderingen ook leiden tot verbetering van de output. Iedere automatisch gegenereerde samenvatting te laten beoordelen door mensen is zeer tijdrovend en daarom is er gezocht naar mogelijkheden voor automatische evaluatie. Een geschikte methode is het maken van een corpus met door mensen gecreëerde samenvattingen. De aanleg van een dergelijk corpus kost één keer inspanning maar kan daarna steeds opnieuw worden gebruikt als vergelijkingsmateriaal. Een belangrijk punt bij deze evaluatiemethode is om een tekst door meerdere mensen te laten samenvatten. Er zijn vaak meerdere manieren om een tekst samen te vatten en uit onderzoek voor zinsextractie blijkt ook dat verschillende annotatoren vaak niet dezelfde zinnen kiezen voor een samenvatting[4]. Er is ook een techniek ontwikkeld om automatisch de kwaliteit van een samenvatting bepalen door te vergelijken met meerdere, door mensen gemaakte samenvattingen, ROUGE [3].

Recente ontwikkelingen in het veld laten ook een verschuiving zien naar meer realistische, gebruikersgestuurde samenvattingen [1]. Het maken van een samenvatting is subjectief: tenslotte bepaalt de vraag van een lezer welk deel van een document belangrijk is. Vraaggestuurd samenvatten helpt daarom ook om de consensus tussen annotatoren te verhogen.

Dit corpus is ontwikkeld voor het evalueren van Nederlandstalige automatische samenvatters. Hierbij hebben we ons vooral gericht op vraaggestuurde multi-document summarization.

3 Het SumNL Corpus

De eerste stap is het maken van clusters van gerelateerde teksten. We hebben 30 vragen gekozen als basis voor de clusters en zoeken handmatig steeds 5 à 25 gerelateerde documenten in een set nieuwsberichten. Het gaat om complexe vragen waarop het antwoord gevonden kan worden in (delen van) meerdere documenten. Een voorbeeldvraag uit het corpus is de volgende: *Nederlandse militairen in Irak in opspraak. Hoe kwamen de Nederlandse militairen in Irak in 2006 in opspraak en wat waren de reacties?.*

De teksten zijn afkomstig van het DCOI corpus (newspapers, periodicals) en nieuwsartikelen van ANP en Novum verzameld als onderdeel van het Daeso monolinguaal parallel corpus. Een karakteristieke eigenschap van de Daeso data is dat een cluster vaak een paar zeer gelijkende artikelen bevat.

Voor ieder cluster zijn door 5 verschillende annotatoren twee samenvattingen (abstracts) gemaakt van verschillende grootte (250 en 100 woorden). Dat levert per cluster 10 abstracts op, en 300 in het totaal. Behalve de samenvattingen in eigen woorden, zijn er ook extracts gemaakt door 10 zinnen uit de krantenartikelen te kiezen.

Daarnaast hebben de annotatoren de zinnen van iedere tekst een score toegekend die hun belangrijkheid aangeeft. Voor iedere zin zijn er dus vijf scores toegekend, deze zijn bij elkaar geplaatst in een tabel, en voor iedere zin is ook de gesommeerde score berekend. Deze belangrijkheidsscores kunnen gebruikt worden om extracts te maken, zowel voor multi-document extracts als voor iedere tekst apart (single-document extracts). Het uiteindelijke evaluatiecorpus bevat dus zowel abstracts als extracts en kan bruikbaar zijn voor de evaluatie van zowel multi-document summarization als single-document summarization.

3.1 Annotatie

Dit corpus is zeer gelijkend op de data sets die beschikbaar worden gesteld door NIST in de Document Understanding Conferences[1]. Het doel van DUC is een evaluatie competitie voor automatische multi-document summarizatie systemen. DUC stelt teksten beschikbaar aan de deelnemers die vervolgens hun systemen automatische samenvattingen laten produceren. Vervolgens evalueert DUC deze samenvattingen door ze te vergelijken met door mensen gemaakte samenvattingen (gold-standard summaries). DUC richt voornamelijk op het Engels.

Bij het maken van de clusters en de samenvattingen zijn zoveel mogelijk dezelfde richtlijnen gebruikt als bij de annotatie van de DUC 2006 data set. Ons corpus heeft een extra toevoeging in vergelijking tot de DUC data, namelijk belangrijkheidsscores voor iedere zin. Voor het toekennen van zinscores konden de annotatoren kiezen uit drie waarden: 0 : zin is niet belangrijk, 1: zin is enigzins belangrijk, 2: zin is heel belangrijk. De uiteindelijke gesommeerde belangrijkheidsscore voor iedere zin is een getal tussen 0 (onbelangrijk) en 10 (alle annotatoren vonden de zin heel belangrijk). De gedetailleerde annotatierichtlijnen zijn te vinden in Bijlage A en B.

[1] voor DUC, zie http://duc.nist.gov

3.2 Structuur

Clusternamen die beginnen met *cluster00* zijn gemaakt op basis van Daeso data, clusternamen beginnend met *cluster20* zijn gebaseerd op DCOI data. De originele nieuwsberichten zijn voor ieder cluster gebundeld in een htmlbestand. Iedere tekst is gescheiden met een lijn. Iedere afzonderlijke tekst begint met een regel waarop de datum van publicatie wordt vermeld en een referentie tussen ronde haakjes naar de oorspronkelijke naam van de tekst zoals deze is opgenomen in DCOI of Daeso. We hebben steeds twee varianten van de htmlbestanden gemaakt. Een zonder toevoegingen, en een waarbij de teksten getokenizeerd zijn met een automatische tokenizer en voorzien zijn van regelnummers. Alle teksten hebben UTF-8 character encodering. De grootte van het corpus is 6.8MB. Het corpus is als volgt gestructureerd. Voor ieder cluster is er een aparte map gemaakt, en iedere clustermap bevat twee mappen: `Data` en `Annotation`. Dit is een voorbeeld van de bestandstructuur van het corpus, in dit geval voor *cluster0001*:

`Data/`	
`cluster0001.topic.txt`	Beschrijving van het cluster topic.
`cluster0001.html`	De originele teksten in html format.
`cluster0001.s1.html`	De getokenizeerde teksten. De teksten zijn gerepresenteerd in html met een zin per regel en iedere zin is voorzien van een zinsnummer.
`Annotation/`	De annotatie is gedaan door 7 annotatoren aangegeven met de letters A t/m G, voor ieder cluster zijn er door 5 verschillende annotatoren samenvattingen gemaakt.
`cluster0001.sentencerankings.txt`	tabel met zinsrankings
Per annotator:	
`cluster0001.A.extr.10s`	extract van 10 zinnen (nummers corresponderen met de htmlfile)
`cluster0001.A.sum.100w`	samenvatting van 100 woorden
`cluster0001.A.sum.250w`	samenvatting van 250 woorden

4 Licenties

Dit corpus is vrij beschikbaar zijn voor niet-commercieel gebruik en zal gedistribueerd worden door de TST-centrale (zie: www.tst.inl.nl/). De artikelen in dit corpus zijn afkomstig van het DCOI corpus en het Daeso corpus waarvan de rechten toekomen aan de TST-centrale. De licentie op de samenvattingen van het SumNL corpus berust bij de Universiteit Antwerpen.

5 Dankbetuiging

De bouw van dit corpus is gefinancïeerd door de Universiteit Antwerpen, Kleine Projecten BOF 2008. Dank aan de studenten die hebben geholpen dit corpus te maken, en dank aan de medewerkers van NIST voor het delen van hun richtlijnen voor DUC 2006.

References

1. H.T. Dang. Overview of duc 2006. In *Proceedings of the Document Understanding Workshop*, pages 1–10, Brooklyn, USA, 2006.
2. Eduard Hovy. Automated text summarization. In Ruslan Mitkov, editor, *Handbook of Computational linguistics*. Oxford University Press, Oxford, UK, 2005.
3. C. Lin.), rouge a package for automatic evaluation of summaries,. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain, 2004.
4. I. Mani. *Automatic Summarization*. John Benjamins, 2001.

A Bijlage: Maken van clusters

Doel: Je start met een vraag of een onderwerp waar je meer over te weten wilt komen. Vervolgens zoek je documenten die deze vraag beantwoorden of die relevant zijn voor het onderwerp.

In het ideale geval zou je met Google in allerlei kranten kunnen zoeken om je set van documenten te vinden. Helaas moet je zoekactie beperkt blijven tot een set van documenten waarvoor we een licentie hebben.

A.1 Het topic

Houd in je achterhoofd dat het uiteindelijke doel van deze taak is: het bouwen van een evaluatiecorpus voor automatische multi-document summarizatie. Je uitgangspunt moet dus steeds een complexe vraag zijn waarop het antwoord gevonden kan worden in (delen van) meerdere documenten. Een vraag over een bepaald feit zoals 'wie heeft de telefoon uitgevonden' en ook vragen om een lijst van feitjes zoals 'welke uitvindingen zijn er gedaan in de jaren 50?' zijn geen goede vragen. Voor voorbeelden van vragen, zie bijlage.

Hier zijn twee criteria waar ieder Cluster en topic aan moet voldoen:

- a) Het antwoord is complex (en kan in ongeveer 100 tot 250 woorden worden samengevat).
- b) Het antwoord kan gevonden worden in de set van relevante documenten.

A.2 Set van relevante documenten

Eisen voor de set van documenten:

- Het minimum aantal documenten voor een topic is 5, streven is 10, en max is 25.
- Ieder relevant document bevat een bijdrage aan het antwoord.
- Redundantie van informatie is te verwachten (verschillende documenten bevatten dezelfde informatie) en is oke.

A.3 Doel

Zoals gezegd is het doel het bouwen van een evaluatiecorpus. Dit documentje beschrijft stap 1, het maken van topics en clusters van relevante documenten. Stap 2 is het schrijven van een samenvatting die een antwoord formuleert op de Topic vraag. Deze handgeschreven samenvatting dienen vervolgens als het evaluatiemateriaal voor automatische summarizatiesoftware.

A.4 Tijd

De geschatte tijd om 1 topic en een set van relevante documenten te maken is 1 uur. Registreer per topic hoeveel tijd het je werkelijk heeft gekost, zodat de tijdsplanning indien nodig kan worden bijgesteld.

A.5 De data

We hebben twee informatiebronnen. De ene heet Daeso, en is een set van parallelle krantenartikelen van twee Nederlandse nieuwsbronnen ANP en Novum van de jaargangen 2006 en 2007. De andere informatiebron zijn de krantenartikelen in het DCOI corpus die de jaargangen 2001 en 2002 beslaan.

B Bijlage: Schrijven van samenvattingen

B.1 Data

Ieder Cluster bestaat uit een topic en een set van teksten. Het topic is een beschrijving van een vraag, een verzoek om informatie over een bepaald onderwerp. De set van teksten is relevant voor deze vraag, en zou het antwoord op de vraag moeten bevatten. Of ruimer gezegd, de documenten geven informatie over het topic onderwerp.

B.2 Lezen

Beginnen met lezen van het Topic. Lees vervolgens de teksten globaal door. Lees de teksten vervolgens in detail door, probeer te bepalen welke stukken van teksten het belangrijkst zijn met betrekking tot het topic.

B.3 Scores toekennen

Ga dan zin voor zin door de tekst en geef iedere zin een score die zijn belangrijkheid uitdrukt:

0: zin is niet belangrijk
1: zin is enigzins belangrijk
2: zin is heel belangrijk

Doe dit voor alle teksten in het cluster.

B.4 Samenvatten

Schrijf vervolgens in je eigen woorden een samenvatting van 250 woorden van de set van teksten. Richt je op het beschrijven van die informatie die betrekking heeft op het topic.

De vorige stap zou je een goed idee van de inhoud van de teksten gegeven moeten hebben. Het is absoluut niet de bedoeling dat je copy-paste gebruikt om de samenvatting te maken. Probeer dus om de samenvatting uit je hoofd te schrijven zonder naar de teksten te kijken. Dit is geen eis, uiteraard mag je best even naar de tekst kijken om iets te verifiëren.

Let er ook op dat je alleen informatie uit de teksten gebruikt en niet je eigen kennis over het onderwerp opneemt in de samenvatting.

Maak vervolgens nog een tweede samenvatting, dit keer een kortere samenvatting van 100 woorden.

B.5 Algemene opmerkingen

Neem alle tijd om de teksten goed en rustig door te lezen. Bij deze opdracht is kwaliteit belangrijker dan kwantiteit. Geef ook aan wanneer je vindt dat een tekst niet erg relevant is voor een bepaald topic.