Statistiska textsammanfattningsmetoder och förbehandling av text

Skribent: Alexander Lemberg
Handledare: Shuha Liu
Tekniska Fakulteten vid Åbo Akademi
Våren 2010

Referat

Problemet med att automatiskt generera en sammanfattningen av en text är att hitta den mest relevanta informationen i originaltexten och presentera den för läsaren. Läsaren bör på basis av sammanfattningen förstå textens centrala innehåll och idéer. I uppsatsen presenteras en generisk statistisk sammanfattningsmetod av text, där samma text förbehandlas på flere olika sätt.

Dessa sammanfattningar ställs i jämförelse med på måfå utvalda meningar ur samma text, och med sig själva. Det finns dock vissa svårigheter med att evaluera vilka utdrag ur en text som är mera informativa. Resultaten ger hur som helst antydningar för vilka förbehandlingsmetoder som lönar sig rent prestandamässigt, samt om referatet över huvudtaget reflekterar över textens centrala innehåll.

Innehållsförteckning

1. Inledning

En sammanfattning är definierad som ”en presentation av materialet i koncentrerad form ur vilken de huvudsakliga idéerna framgår” (http://www.thefreedictionary.com). Att skapa sammandrag av text är möjligt eftersom dokument oftast innehåller text som inte är viktig för att förstå de centrala idéerna. Dessutom brukar dokument ha redundant information, då viktiga fakta oftast presenteras flera gånger i texten [4]. Ett problem är dock att genererade sammanfattningarna inte alltid representerar textens centrala innehåll [1].

1.1 Utdrag och Referat

Sammanfattningsmetoder kan klassificeras som *referat* (abstract på engelska), vilken är en av författaren egenhändigt formulerad text som beskriver det väsentligaste innehållet i en text i nya ord, eller som en sammanfattning som baserar sig på *utdrag* ut texten (extract på engelska). Ett *utdrag* ur texten är alltså ett plock av de viktigaste meningarna från originaltexten, vanligtvis presenterade i den ordning som de förekommer i originalet.

Då man konstruerar en sammanfattning av typen *referat* behövs förståelse av texten samt en språkvetenskaplig analys för att kunna hitta nya uttryck för att generera en ny kortare text som uttrycker det viktigaste innehållet i originaltexten [1].

Även om att skapa ett självständigt referat vore det bästa sättet att skapa en sammanfattning på (detta är alltså så som vi människor gör), så är teknologin för de språkvetenskapliga metoderna ännu tyvärr så pass outvecklad att sådana metoder i praktiken inte är utförbara [1]. P.g.a. detta fokuseras det i uppsatsen på sammanfattningsmetoder för att skapa - *utdrag* - ur originaltexten, och kommer i fortsättningen att syfta på utdrag av meningar ur text då jag pratar om sammanfattningar eller sammandrag.

Ett utdrag är, i kontrast till ett referat, en presentation av texten där de flesta meningar är bortlämnade. Endast de viktigaste meningarna presenteras för läsaren, oftast i omanipulerad form och i samma ordning som de presenterats i originaltexten[4].

Utmaningen med att skapa utdrag av en text är att hitta de mest relevanta meningarna i originaldokumentet eller dokumenten som beskriver texten sakliga innehåll.

Metoder för att göra utdrag ur texten är vanligtvis uppbyggda i flere steg, där man i varje steg kan ha många olika parametrar och alternativ. Ofta gör man statistiska undersökningar av texten, och på basen av dem kan välja vilka delar av texten som skall tas med i utdraget. Fördelen med utdrag istället för sammandrag är att det är en robust metod som p.g.a. dess statistiska analyser även är oberoende av textens språk. Nackdelen är att meningarna i ett utdrag oftast inte hänger så bra ihop.

1.2 Riktade sammanfattningar

Att ta i beaktande vad användaren är ute efter kan vara viktigt då man gör sammanfattningar. Detta har t.ex. Google gjort med sina sökord. Det finns även andra sätt att utföra detta på, t.ex. att bara sammanfatta meningar som handlar om ett visst ämne, eller t.ex. meningar som innehåller mycket siffror, ifall det är det som läsaren är ute efter. Dessa kallas för *användar-fokuserade* sammanfattningar. I den här uppsatsen kommer jag dock inte att ta detta i beaktande, utan kommer att använda metoderna för att skapa *generiska* sammanfattningar. En generisk sammanfattning ger en generell helhetsbild över dokumentets eller dokumentens innehåll [5].

1.3 En ideell sammanfattning

Oftast då en studerande får en 1000 sidors bok att läsa är tiden för knapp för att den studerande skall hinna läsa igenom boken från pärm till pärm, även om tentamen kommer att omfatta hela bokens innehåll. Detta resulterar ofta i att den studerande kollar i gamla tentamina efter vilka stycken det lönar sig att läsa, läser dem, skriver en godkänd tentamen, förmodligen utan att ha så god kontroll över ämnet som tentamensresultatet framhäver.

Om man kunde (elektroniskt) plocka ut endast de viktigaste styckena och meningarna med tabeller och grafer ur boken med en sammanfattningsalgoritm, presentera dem på ett tiotals sidor och dessutom bevisa att sammandraget tar upp varje viktigt ämne ur texten, vore den informationen guld värd för en studerande.

1.4 Fokusering

I uppsatsen kommer en och samma statistiska algoritm att användas för att sammanfatta en mängd av text dokument och presentera det centrala innehållet från alla dokument tillsammans. Texten kommer dock att förbehandlas på olika sätt, och resultaten ställs sedan i jämförelse med en metod som på måfå plockar ut 12 st. meningar ur texten. Ett problem är här även att bedöma resultatet, eftersom det ”bästa” resultatet är beroende av användaren och sammanhanget. En expert inom området kan få mera ut av sammandraget i en facktext än av vilka utplockade meningar som helst från samma text.

2. Förbehandling av texten

2.1. Teckenuppdelning

För att kunna analysera och hantera texter bör vi först göra en s.k. förbehandling av texten där vi delar upp texten i mindre enheter. Oftast delar man upp texten i t.ex. ord, meningar, paragrafer och dokument. För att dela upp dokumenten i dessa mindre enheter använder jag mig i mina testförsök av ”Punkt tokenizer” från Kiss & Strunk [3] för att dela upp texten i mindre enheter.

Att dela upp texten i meningar är ett viktigt grundjobb som bör göras så noggrant som möjligt. Om meningarna inte är uppdelade rätt kan det ha stor inverkan på slutresultatet. Om flere korta meningar av någon orsak slås ihop (t.ex. om en mening slutar med en förkortning) kan de få högre poäng än om de vore vara presenterade i sina korrekta längder. ”Punkt tokenizer” från Kiss & Strunk har dock visat sig klara av denna uppgift hyfsat bra. Endast rubriker som slutar utan punkt har i experimenten observerats att slås ihop med den följande meningen i texten.

2.2. Normalisering av tecken

Efter att alla meningar och ord är uppdelade kan man normalisera alla tecken till små bokstäver. Bokstäver som har stor begynnelsebokstav ersätts av samma ord med små bokstäver. Man bör avgöra ifall t.ex. orden ”Picture” och ”picture” skall tolkas som samma ord då statistik räknas över texten. Ibland kan detta hjälpa, och ibland inte, vilket även kommer fram i resultatet av uppsatsen.

Ett sätt att avgöra ifall begynnelsebokstaven skall normaliseras eller ej, kunde vara att ta reda på ifall ordet är ett namn på någonting, och i så fall låta begynnelsebokstaven stå kvar i sin ursprungliga form. Då skulle statistiken över orden klara av att åtskilja ord som ”George Bush” och ”a bush”. Till detta behövs dock information om vilka ord som är namn och inte, och detta kommer inte heller att vidare beaktas i denna uppsats.

2.3. Ordstammar

Till förbehandlingen av texten hör även ofta att generera stammarna av orden, dvs. man normaliserar orden. Detta betyder således att t.ex. orden ”connections”, ”connection” och ”connected” alla representeras av sin stam ”connect”. Ofta förbättras resultaten för IR (Information Retrieval) system då man tar bort ändelserna från alla ord och representerar dem med stammarna. Detta minskar även på antalet termer i texten, och reducerar således till viss mån storleken och komplexiteten för IR system [7].

För att generera stammar av orden används Porter’s Stemmer algoritm för engelska [7] (eftersom algoritmen är gjord för det engelska språket, kommer engelska exempel att användas). Då man slopar ett ords suffix i IR system är det vanligtvis för att förbättra utförandet och prestandan i processen. Detta betyder att det inte alltid är självklart hur stammen kommer att se ut, även om vi vet hur den språkvetenskapligt borde se ut.

Låt oss välja två ord O1 och O2. Det bästa sättet att avgöra när det är giltigt att ta bort suffix från orden O1 och O2 för att producera en gemensam stam S, är om det inte visar sig vara någon skillnad mellan påståendena ’this document is about O1’ och ’this document is about O2’. Ifall O1 skulle vara = ”connection”, och O2 = ”connections” verkar det vara ganska rimligt att representera dessa två ord med en gemensam stam.

Ifall O1 vore = ”relate” och O2 = ”relativity” verkar det å andra sidan inte vara motiverat att representera dessa med en gemensam stam, i synnerhet ifall dokumenten handlar om teoretisk fysik. För de flesta fall av O1 och O2 kommer det att finnas olika åsikter om de borde representeras av en stam eller ej, på samma sätt som att avgöra om ett nyckelord är av relevans till någon text [7].

2.4. Ordval

Då texten är uppdelad i ord och meningar kan ord som inte bidrar med någon saklig information till texten filtreras bort. Dessa s.k. ”stop-words” som inte ger någon saklig information är t.ex. ”and”, ”the”, ”when”, ”that”, ”who” m.fl. Man bör dock veta i vilka fall man kan filtrera sådana ord, och när man inte kan göra det. Om man t.ex. vill göra en användar-fokuserad sammanfattning med sökord kan det gå dåligt om användaren söker på t.ex. ”The who” eller ”Take that”, eller andra fraser som består av stop-words. Eftersom denna uppsats fokuserar på generiska sammanfattningar, så kommer jag att filtrera bort dessa stop-words, med ett litet undantag i fras-sökningar (*n*-grams) där vi lämnar kvar stop-words, men däremot tar bort alla fraser som består av endast stop-words. Denna procedur beskrivs vidare i kapitel 3.2.2.

3. Statistiska metoder

De typiska sätten att ta itu med automatisk textsammanfattning består av att välja ord, poängsätta ord, poängsätta meningar, och sedan välja ut de meningarna som fått högsta poäng [1]. Sätten att poängsätta ord och meningar är vad som skiljer alla dessa metoder från varandra.

3.1. Poängsättning av ord

3.1.1. TF

Term Frequency (TF) poängsätter ord enligt hur ofta de förekommer i texten. Ju oftare ett ord förekommer, desto högre blir dess poäng. Då vi sammanfattar flere dokument dividerar vi ett ords TF poäng med antalet dokument som texten består av.

$TF_j = f_j$ / |D| , där f_j är frekvensen av termen j i alla dokument, och |D| är det totala antalet dokument i texten.

Om vi har 2 dokument där ordet ”picture” förekommer 11 gånger, blir TF poängen för picture således 5.5 poäng. Denna metod är föreslagen i [6].

3.1.2. IDF

Inverse Document Frequency (IDF) är en metod för att räkna ut ett ords allmänna betydelse i en samling av dokument. Nackdelen med TF är att, då ett ord finns i så gott som alla dokument är TF oanvändbart för att särskilja på relevanta dokument [1]. T.ex. finns ”stop-words” ”and” och ”the” garanterat i alla dokument, och har högt TF poäng. IDF straffar dock ord som förekommer i alla dokument:

$$idf_j = \log (|D| / |\{d : t_j \in d\}|) ,$$

där |D| är totala antalet dokument och $|\{d : t_j \in d\}|$ är antalet dokument som innehåller ordet t_j.

Om antalet dokument är 5, och ordet ”picture” finns i 2 av dokumenten, blir $idf_{picture}$ = log (5 / 2) = 0.398

Detta betyder att *IDF* värdet för ett ord blir högt ifall det förekommer i endast ett fåtal dokument, medan *IDF* värdet blir nära 0 ifall det förekommer i nästan alla dokument.

Eftersom experimentet i denna uppsats består av endast ett dokument, representeras texten som en samling av meningar vid IDF uträkningen, istället för en samling av dokument. Detta var föreslaget i [1].

3.1.3. TF*IDF

Problemet med IDF är att det inte är möjligt att åtskilja olika dokument som råkar ha samma ordförråd, fastän vissa ord kan förekomma oftare i den ena dokumentet [1]. För att åtgärda detta multipliceras IDF poängen med TF poängen enligt följande:

$$tf\text{-}idf_j = TF_j * IDF_j$$

Om antalet dokument är 5, och ordet ”picture” finns i 2 av dokumenten och har *TF* poäng 5.5, blir då $tf\text{-}idf_{picture}$ = 5,5 * 0,398 = 2,189.

3.2. Poängsättning av meningar

3.2.1. Centroid metoden

En centroid är en mängd av ord som statistiskt representerar ett eller flere dokument. Ett dokument är representerat som en viktad vektor av TF-IDF poäng. Centroiden är den mängd av ord som har TF-IDF poäng över ett visst tröskelvärde.

Centroid metodens ultimata mål är att tilldela ett poäng till varje mening i alla dokument. En menings S_i centroid-poäng C_i är summan av alla centroid värden $C_{w,i}$ av alla ord i meningen [6].

$$C_i = \sum_w C_{w,i}$$

Poängen för meningarna normaliseras slutligen till ett värde mellan 0 och 1. Således kommer meningen/de meningar med högsta Centroid poäng att få poängen 1.0, och de övriga meningarnas poäng är 0.0 och < 1.0.

3.2.2. *N*-gram

Ett *n*-gram är en sekvens av *n* stycken ord i texten. *N*-gram som förekommer flere gånger i texten är oftast av intresse. T.ex. kunde *3*-grammet ”President George Bush” vara ett intressant *n*-gram om det förekommer ofta i texten [8].

Definitionen på *n*-gram beror dels på vad man anser att är ett ord. T.ex. kan man överväga ifall ”Herr Björk” och ”en björk” skall representera samma ord (björk) eller inte. Detta problem ligger inom området ”igenkännande av en namngiven entitet” (named-entity recognition) inom forskning i naturliga språk, och kan användas som hjälp för att sammanfatta texter. Jag kommer dock inte att använda någon sådan metod i dessa försök. Man bör även klargöra ifall ord med samma stam skall representera samma ord (fråga, frågade, frågar), eller om t.o.m. ord med samma betydelse skall representera samma ord (tycka, anse) [1].

I de olika försöken kommer jag att använda *N*-grams av både ordstammar och originalorden. *N*-gram som enbart innehåller ”stop-words” ignoreras i bägge fall.

Målet med dessa *n*-grams, lika som med centroids, är att ge poäng åt meningar i texten. Detta görs genom att först räkna ut TF*IDF för varje *n*-gram i texten på samma sätt som för varje ord i texten:

$tf\text{-}idf_j = TF_j * IDF_j$, där TF_j är frekvensen av *n*-grammet *j* i alla dokument, och IDF_j är IDF värdet för n-grammet *j*.

N-gram som har sin TF-IDF poäng över ett visst tröskelvärde beaktas, och de övriga *n*-grammen förkastas.

Om en mening innehåller ett *n*-gram vars *TF-IDF* poäng är över tröskelvärdet, kommer meningen att tilldelas *n*-grammets *TF-IDF* värde som poäng, multiplicerat med antalet gånger n-grammet förekommer i meningen. Poängen för meningarna normaliseras slutgiltigen till ett värde mellan 0 och 1. Således kommer meningen/de meningar med högsta *n*-gram poäng att få poänget 1.0, och de övriga meningarnas poäng är 0.0 och < 1.0.

3.2.3. Positions poäng

De första meningarna i en text har en tendens att reflektera över innehållet ganska bra [4]. Positions poäng för meningar tar inte meningarnas innehåll i beaktande, utan ger endast poäng åt i vilken position av dokumentet de råkar befinna sig, och räknas med funktionen $\sqrt{1/i}$, där *i* är ett sekventiellt nummer startandes från 1 [4].

Detta betyder i praktiken att meningarnas positionspoäng är (1, 0.7071, 0.5773, 0.5000 osv…).

3.2.4. Meningarnas längd

Ofta vill man inte ha allt för korta meningar med i ett sammandrag, även om de skulle vara högt poängsatta. Då meningarnas längd tas i beaktande, lämnar man oftast bort meningar ur sammandraget som inte är tillräckligt långa. Detta sker i praktiken genom att en mening som inte uppfyller längdkriteriet får poängen 0 fastän den skulle ha en hög centroid- eller positionspoäng.

4. Sammanställning av meningarnas poäng

Meningarnas centroid-, n-gram- och positionspoäng räknas ihop med vikter som definieras av användaren. Varje poängsättningsmetod får en egen vikt som bestämmer hur mycket den kommer att influera resultatet av sammandraget. Om vikterna exempelvis för en menings s_i centroid och *n*-gram poäng är 2.0 respektive 0.5, och låt oss säga att centroid och *n*-gram poäng för meningen är 0.7 respektive 1.0. Totala poängen *C* för meningen s_i blir således 2.0*0.7 + 0.5*1.0 = 1.9 med de givna parametrarna.

Ett undantag finns här då vi även beaktar längden på meningarna, så som beskrivet i stycke 3.2.4. Då tillämpas inte vikter för dem i poängräkningen, utan meningar som inte uppfyller längdkravet får direkt 0 poäng som diskuterades tidigare.

5. Poängjustering av meningar

Problemet med de poäng vi räknat ihop hittills, är att vi kan som resultat få meningar som påminner mycket om varandra, eller t.o.m. flere likadana meningar ifall sådana finns i texten. Ett sätt att ta itu med detta problem är att jämföra liketer i meningar och antingen ta bort meningar eller ändra poäng för meningar som påminner om varandra. Detta görs med poängjusteringsmetoder, och metoder för att beräkna likhet hos meningar som beskrivs i detta kapitel.

5.1. Likhet hos meningar

För att kalkylera likhet mellan meningar representerar vi, som föreslås i [6] meningarna som vägda vektorer av *IDF* poäng och räknar sedan ut vinkeln mellan dessa två vektorer. Eftersom vektorernas vikter inte kan vara negativa i det här fallet, kommer vinkeln mellan dem aldrig att bli mer än 90°. Likheten mellan vektorerna A och B mäts alltså av *IDF* värdet av överlappande ord, och definieras som:

$$\mathrm{Sim}(A, B) = \cos(\theta) = \frac{A \cdot B}{||A|| ||B||} \qquad \text{Formel 5.1.1}$$

, där ||A|| och ||B|| är normen av vektor A och B. Två meningar som sammanfaller, dvs. har $\cos(\theta) = 1$ definieras som lika, och två meningar som har $\cos(\theta) = 0$ klassas som

olika. Detta betyder även att vektorerna som jamförs måste vara av samma dimension, vilket löses med att fylla ut icke-överlappande ord med nollor.

5.1.1. Uträkning av likhet:

Antag att vi har två meningar *A* och *B* med orden [a, b, c, d, a, b] respektive [e, b, a, c, e], där alla unika ord har ett *IDF* värde. Dimensionen på dessa vektorer är 6 respektive 5, och är således olika. Meningarna *A* och *B* slås ihop till en mängd *M* av ord [a, b, c, d, e], och alla duplikat av ord blir således eliminerade i *M*. Listan *M* av unika ord gås igenom för vardera meningen, där frekvensen för varje ord i meningarna *A* och *B* som finns i *M* räknas ut, och bägge meningar får således varsin vektor A_f och B_f där frekvensen för varje ord som finns i *M* framgår. Bägge vektorer A_f och B_f har samma dimension $|M|$, och samma index för orden i *M*.

Vektorerna A_f och B_f ser nu ut som [2, 2, 1, 1, 0] respektive [1, 1, 1, 0, 2]. Nu söks *IDF* värdet för varje ord i *M* från tidigare kalkylationer (stycke 3.1.2), och multipliceras med frekvensen för ordet i meningarna.

Låt oss säga att *IDF* värdet för orden [a, b, c, d, e] är [1.0, 2.23, 0.5, 5.3, 2.0]. Vektorerna A_f och B_f blir då [2.0, 4.46, 0.5, 5.3, 0.0] respektive [1.0, 2.23, 0.5, 0.0, 4.0]

Nu då meningarna representeras av vektorer av *IDF* poäng, där bägge vektorer har samma dimension, kan formel 5.1.1 användas för att räkna ut cosinus likheten för vektorerna [10]. Cosinus likheten mellan vektor *A* och *B* blir således ca 12.1958 / 34.07 ≈ 0.36.

Eftersom cosinus likheten mellan två *IDF* vektorer enbart bestäms av IDF vektorernas element, är det viktigt att IDF poänguträkningen är pålitlig. Skall IDF räknas t.ex. på ord stammarna eller på själva orden? Skall ord med samma betydelse få samma IDF poäng? Skall man skilja på samma ord med lite eller stor begynnelsebokstav?

5.2. MMR poängjustering

MMR ”reranker” (eng. Maximal Marginal Relevance) är en poängjusterare vars ultimata mål är att eliminera redundanta meningar från textsammanfattningen, medan relevanta meningar bör bibehållas [9].

MMR börjar med att hitta den mening s_i med högsta poäng som ännu inte justerats, och mäter cosinus likheten mellan alla redan MMR-justerade meningar. Den redan justerade meningen s_j som har högsta cosinus likhet med mening s_i blir sedan vald. Efter det används följande formel för att räkna ut ett nytt poäng åt mening s_i:

$$s'_i = \lambda Cs_i - (1-\lambda) * \mathrm{Sim}(s_i, s_j) \quad , \qquad \text{Formel 5.2.1}$$

där λ är en användardefinierad vikt mellan 0 och 1, och Cs_i är ursprungliga poäng för mening s_i och s'_i är en ny poäng för mening s_i. Om λ är 1 hålls meningarna kvar med sina ursprungliga poäng, och λ läggs nära nolla baserar sig poängen mest på olikhet mellan meningarna.

5.3. Cosinus poängjustering

Cosinus poängjusteringen försöker se till att två likadana meningar inte kommer med i sammanfattningen. Algoritmen går systematiskt genom alla meningar *s*, ordnade enligt högsta poäng, och jämför cosinus likheten mellan de meningar med högre poäng som inte blivit eliminerade. Då cosinus likheten mellan två meningar s_i och s_j räknat enligt formel 5.1.1 överstiger ett visst gränsvärde, elimineras mening s_j, där s_j är den mening som ännu inte lagt med till resultatet, och s_i är en mening som inte blivit eliminerad och har högre poäng än s_j.

På det här sättet kan vi vara säkra på att inga två meningar som har en cosinus likhet över exempelvis värdet 0.8 kommer med i sammanfattningen.

5.4. Sammanfattning

Både cosinus poängjusterarinsalgoritmen och MMR poängjusteraringsalgoritmen räknar ut vinkeln mellan två vektorer för att bestämma likheten mellan två meningar, där meningarna är representerade som varsin vektor av IDF värden. Största skillnaden i metoderna är att cosinus poängjusterarinsalgoritmen tar bort meningar där cosinus för vinkeln mellan dem ligger över ett visst tröskelvärde, medan MMR poängjusterarinsalgoritmen justerar alla meningars poäng på basis av cosinus för vinkeln mellan dem.

Att omsorgsfullt förbehandla texten då IDF uträknas är viktigt, eftersom ordens IDF värden slutligen kommer att bestämma vilka meningar som tas med i sammanfattningen.

6. De olika försöksmetoderna

6.1. En algoritm för att generera sammandrag

I algoritmen varieras mellan att eliminera ”stop-words”, lämna kvar stop-words, och att reducera orden till små bokstäver samt låta alla versaler stå kvar. Efter det skapas ännu en variation genom att generera stammar av orden med Porters stemming algoritm [7], och genom att inte göra det. I alla försök kommer dock icke-litterära tecken såsom ”,% ()” samt siffror att filtreras bort då ordens poäng och meningarnas likhet räknas. Denna förbehandling kommer att ge olika resultat för bl.a. ordens eller stammarnas *IDF* poäng. Sedan räknas TF-IDF poäng för de ord eller stammar som finns kvar, och en gemensam centroid genereras för alla dokument som skall sammanfattas. *N*-gram statistik räknas sedan ut enligt metoden som föreslagits i 3.2.2.

Meningar som är kortare än tröskelvärdet får poängen 0, sedan räknas poängen ihop med godtyckliga vikter enligt metoden i kapitel 4. Cosinus poängjusterarinsalgoritmen eliminerar sedan meningar som påminner mycket om varandra, och MMR poängjusterarinsalgoritmen justerar poängen enligt hur lika de kvarvarande meningarna är. Sedan väljs de 12 mest representativa meningarna ut, dvs. dem med högsta poäng, och inkluderas i sammandraget.

- Förbehandling: eliminera stop-words, normalisera till små bokstäver och tillämpa Porter stemming / tillämpa inte Porter stemming,
- Poängsättning av stammar/ord: TF-IDF
- Poängsättning av meningar: centroid med tröskelvärde 0.12
- Poängsättning av meningar: *n*-grams av stammarna/orden, i det här fallet 3-grams. Vi väljer ut de 30 första 3-grams av högsta poäng för att poängsätta meningarna enligt 3.2.2.

- Längdjustering: meningar med färre än 7 ord förkastas, samt meningar med mera än 190 ord.
- Ihopräkning av centroid, n-gram och positions poäng för meningar med vikterna centroid: 1.0, n-gram: 1.0, position: 0.2
- Poängjustering med cosinus poängjusterarinsalgoritm: eliminera liknande meningar med tröskelvärdet 0.70
- Poängjustering med MMR poängjusterarinsalgoritm: återge meningarnas poäng på basen av deras cosinus likhet, med λ definierat som 0.6

Alla dessa parametrar är godtyckligt valda med sunt förnuft. Om jag varierade på dessa parametrar i den här uppsatsen skulle alla olika resultat inte rymmas med på pappret. Även den enskilda algoritmerna och ordningen för dem är valda med sunt förnuft och implementerade i Python och biblioteket NLTK (http://www.nltk.org/ ,version 2.0b8). Man kunde givetvis kasta om ordningen på vissa algoritmer, som att t.ex. köra centroid poängsättningen före n-gram poängsättningen el.dyl.

Resultatet som fås med från denna algoritm, med och utan Porter stemming, ställs i jämförelse till 12 st. på måfå utplockade meningar. Med alla variationer på denna algoritm i förbehandlingen, kommer totalt 8 st. olika sammanfattningar av samma text att produceras.

7. Resultat

Texten som sammanfattas är en promemoria ”Gene Technology Strategy and Action Programme of the Ministry of Agriculture and Forestry 2009-2013” från skogs- och jordbruksministeriets hemsida. Texten är i ett dokument, och består av 86kb oformatterad text då dokumentets egna sammandrag och källförteckning tagits bort.

7.1. Alla meningar

Här presenteras mängden av alla meningar som som kommit ur de olika testförsöken, tillsammans med vilka dokument de kommit ifrån, samt ordningstalet för utvalda meningen ur dokumentet. Meningarna är presenterade i tabellen enligt deras inbördes

ordningsnummer för respektive dokument. Denna tabell kommer att hänvisas till då resultaten ur de olika försöken jämförs. Tabell 7.1.1. består av 16 olika meningar, vilket betyder att försöksmetoderna har plockat ut väldigt likadana meningar. I tabell 7.1.1. presenteras inte de meningar som plockats ut på måfå för jämförelse.

Mening nr.	Text
27	The usefulness of gene technology in research and product development is likely to be proven in future by the speed and accuracy of gene technology methods and the new knowledge that these methods can provide.
62	In this Strategy: - customer means a consumer, producer or other customer - gene technology means a group of methods used to isolate, analyse and transfer genes on a molecular level1 (gene technology includes, for example, gene transfer, determining the order of DNA bases, i.e sequencing, the use of DNA markers in selection, production of genetically modified organisms and gene therapy) - quality means hygienic, nutritional, sensory, technical and ethical quality as well as environmental and service quality.
194	The use of gene technology in the areas of industry and commerce administered by the Ministry of Agriculture and Forestry is currently one of the Ministry's most strictly regulated activities.
199	The most important of these in the administrative sector of the Ministry of Agriculture and Forestry are the regulation on genetically modified food and feed, regulation concerning the traceability and labelling of genetically modified organisms and the traceability of food and feed products produced from genetically modified organisms, directives on the open and contained use of genetically modified organisms, seed trade directives, directive on the marketing of forest reproductive material, directive on additives in feedingstuffs, directive on plant protection products, regulation on the transboundary movement of genetically modified organisms, as well as regulation on organic production of agricultural products.
200	The starting point in the legislation is that the use of genetically modified organisms is based on a prior approval system, founded on scientific risk assessment and the precautionary principle, labelling of genetically modified products as well as risk management with adequate monitoring and with general or case-specific follow-up.
203	An important objective in the control of the use of genetically modified organisms is the advance prevention of the adverse impacts of the use of gene technology on the environment, health and property.
205	Provisions applicable to (environmental) damages caused by the use of gene technology have been laid down at least in the Gene Technology Act

	(377/1995), the Act on Compensation for Environmental Damage (737/1994), the Product Liability Act (694/1990) and the Tort Liability Act (412/1974).
211	In Finland, provisions on the approval and use of genetically modified organisms are laid down in the Gene Technology Act, Animal Welfare Act, Seed Trade Act, Plant Protection Products Act and the Act on Trade in Forest Reproductive Material.
221	For implementation of the tasks under the Gene Technology Act, the Government appoints the Board for Gene Technology (GTLK) under the Ministry of Social Affairs and Health to sit for five-year terms, with representatives from the Ministry of Employment and the Economy, the Ministry of Agriculture and Forestry, the Ministry of Social Affairs and Health and the Ministry of the Environment.
224	These include the Finnish Food Safety Authority (Evira), National Institute for Health and Welfare (THL), National Agency for Medicines, MTT Agrifood Research Finland, Finnish Forest Research Institute (Metla), Finnish Game and Fisheries Research Institute (RKTL), National Supervisory Authority for Welfare and Health Valvira, Finnish Environment Institute (SYKE), Finnish Institute of Occupational Health and the Technical Research Centre of Finland (VTT).
278	In carrying out its tasks, the Ministry of Agriculture and Forestry and its entire administrative sector take an official stand regarding gene technology and its use and on the approval and control of products manufactured by means of this technology in agriculture and the food sector in various contexts, including the planning of policies, preparation of legislation (international, EU and national), inspection and control operations and research.
284	In addition, a cross-administrative ad-hoc working group on the control of the use of gene technology also operates under the Ministry of Agriculture and Forestry's Department of Food and Health.
285	Represented in the group are the principal officials and authorities dealing with questions related to the control of the use of gene technology from the Ministry's administrative sector, from the Board for Gene Technology and from the environmental administration.
289	In accordance with Government Decree 910/2004, as amended by Government Decree 135/2008, the national position in decision making regarding application for a permit relating to genetically modified foods and feeds will be coordinated in such a way that the responsibility for foods and feeds rests with the Ministry of Agriculture and Forestry and the Gene Technology Board carries the responsibility for questions related to environmental safety.

326	5) Under the Ministry's guidance it will be ensured that Evira will keep the control strategy for the use of gene technology up-to-date and that Evira's cooperation group for gene technology will draw up, yearly, a coordinated control programme for genetically modified seeds, feeds and foods.
336	The Ministry of Agriculture and Forestry, in cooperation with the Academy of Finland, Tekes and international financiers, seeks to develop forms of financing for biotechnology and gene technology research.
338	The Environment and Natural Resources Consortium established by the Ministry of Agriculture and the Ministry of the Environment tightens the cooperation between the research institutes of the administrative sectors and provides opportunities to boost the efficiency of research also in questions related to gene technology.
367	- selection, in cooperation with researchers and plant breeders, of the research subjects, i.e crop plants and varieties as well as the traits, that from the viewpoint of Finland's plant production are important in the following priority areas: - adaptation to climate change - improved energy use of biomass - improved nutritional quality of food products - promotion of coordinated cooperation between researchers and plant breeders in the key areas mentioned above, on both the national and international level.
371	National Advisory Board for Biotechnology (BTNK), Boreal Plant Breeding Ltd, Board for Gene Technology (GTLK), higher education institutes, Central Union of Agricultural Producers and Forest Owners (MTK), agricultural advisory organisations, National Agency for Medicines, Potato Research Institute (Petla), Sugar Beet Research Centre, Ministry of Social Affairs and Health, Ministry of Employment and the Economy, industry and commerce, Customs Laboratory, Novel Food Board UELK, Ministry of the Environment.
529	Consumers must play an active part in the discussion on the utilisation of GMOs in the administrative sector of the Ministry of Agriculture and Forestry and be able to follow the research and approval of products for marketing, for example, via the websites of the National Advisory Board on Biology, the Board for Gene Technology and Evira.
534	- labelling relating to the production methods of foodstuffs and feeds or the provision of information on these issues by means of other systems is to be improved on the basis of feedback received from consumers - active participation in the discussion on the utilisation of gene technology, and the development of communication regarding GM products and the activities of the administration and food chain relating to these - active participation in the development of preconditions and systems for consumer information at the Community level.

535	- encouraging researchers to participate in the public debate on gene technology and to present real, scientifically proven advantages and disadvantages of the application of gene technology in agriculture and the food industry, and implications for consumers and the environment - scientific consumer research will be conducted to gauge consumer opinion regarding gene technology and to determine how consumer attitudes can be influenced and how consumer information can be improved.
559	Ministry of Agriculture and Forestry, MTT Agrifood Research Finland, Finnish Food Safety Authority (Evira), Finnish Forest Research Institute (Metla), Finnish Game and Fisheries Research Institute (RKTL).

Tabell 7.1.1

7.1.1. Metod 1

Förbehandling: endast teckenuppdelning.

7.1.2. Metod 2

Förbehandling: teckenuppdelning, eliminering av ”stop-words”.

7.1.3. Metod 3

Förbehandling: teckenuppdelning, samt Porter stemming algoritm.

7.1.4. Metod 4

Förbehandling: teckenuppdelning, eliminering av ”stop-words” samt Porter stemming algoritm.

7.1.5. Metod 5

Förbehandling: teckenuppdelning, och konvertering till små bokstäver.

7.1.6. Metod 6

Förbehandling: teckenuppdelning, konvertering till små bokstäver, eliminering av ”stop-words”.

7.1.7. Metod 7

Förbehandling: teckenuppdelning, små bokstäver, Porter Stemming algoritm.

7.1.8. Metod 8

Förbehandling: teckenuppdelning, små bokstäver, Porter Stemming algoritm samt eliminering av ”stop-words”.

7.1.9. Resultat från metod 1 – 8

Mening nr.	Metod 1	Metod 2	Metod 3	Metod 4	Metod 5	Metod 6	Metod 7	Metod 8
27				x				x
62	x	x	x	x	x	x	x	x
194					x		x	
199	x	x	x	x	x	x	x	x
200	x	x	x			x		
203	x	x	x	x	x	x	x	
205							x	
211 *)		x		x		x		x
221	x	x	x	x	x	x	x	x
224 *)		x		x		x		x
278	x	x	x	x	x	x	x	x
284	x	x	x		x	x	x	
285					x			
289			x	x	x		x	x
326	x							
336		x		x		x		x
338	x		x		x		x	
367			x	x				x

371 *)		x		x		x		x
529	x	x	x		x	x	x	x
534			x				x	
535	x				x			
559	x							

Tabell 7.1.2

De celler i tabell 7.1.2 som har ett x, indikerar att metoden för den kolonn där x befinner sig i innehåller meningen med den nummer som första kolonnen i samma rad har. Således har t.ex. Metod nr. 1 meningarna 62, 199, 200, 203, 221, 278, 284, 326, 338, 529, 535 och 559 som kan läsas ur tabell 7.1.1. Raderna i tabell 7.1.2 som är indikerade med *) representerar sådana meningar i utdraget som – antagligen – inte bidrar med någon saklig information från texten. Dessa meningar är uppräkningar vars betydelse är svår att förstå utan ett sammanhang.

7.2. Jämförelse

Som kan ses ur tabellen 7.1.2 påminner sammanfattningarna om varandra. Att sedan avgöra vilken sammanfattning som är mest informativ är ett subjektivt avgörande. Alla metoder producerar 4 meningar som är precis likadana. Vi kan även se att metod, 2, 4, 6, 8, dvs. de metoder som filterar bort ”stop-words” inkluderar meningarna nr. 211, 224 och 371 som har konstaterats att inte bidra med någon saklig information ur texten. Vad som även skiljer dem åt som inte framgår ur tabell 7.1.2, är prestandan som visas i tabell 7.3.2.

För att även ställa dessa metoder 1-8 i jämförelse tar vi 12 st. meningar som valts på måfå från texten:

Mening nr.	**Text**
372	Animal nutrition.
92	By 2007-2008, 615 approvals for use as food, feed or in cultivation - representing a total of 129 modifications of 23 different crop plants - had been

	granted in 55 countries.
381	- support for gene technology research which aims to develop new, more rapid, accurate and economical methods to ensure the safety, quality and genuineness of feedstuffs and for the control needs.
295	In addition, a feedstuffs sub-committee established by the Ministry will be heard on the decision.
335	Several strategic concentrations of excellence industry have been established in Finland, where gene- and biotechnological research is conducted in close cooperation with the industrial applications.
229	Utilisation of gene technology imposes new demands on the control systems.
334	Background.
366	Measures to be taken.
175	Judging by the contents of their shopping bags, the majority of the many respondents who stated that they would not buy genetically modified products did not, in practise, actively avoid foodstuffs labelled as genetically modified.
44	The health impacts of foods can be improved, even in the primary production phase, by refining their quality characteristics through the application of genetic know-how, thus broadening the existing range of functional foods.
131	The gene insert site in the plant can, in some cases, be selected beforehand.
421	Of the forests in economic use in Finland, 95% have been certified in accordance with the FFCS.

Tabell 7.2.1

I tabell 7.2.1 finns 12 på måfå utplockade meningar ur texten. Eftersom alla jämförelser är så gott som subjektiva tas det inte någon direkt ställning till vilka meningar som är bättre, dvs. meningarna från metod 1-8 eller meningarna i tabell 7.2.1.

I tabell 7.2.1. hittas åtminstone 4 meningar som inte ger någon substans av innehållet i texten, närmare sagt meningarna 372, 334, 366 och 131. Ur metoderna 1-8 hittades endast två st. sådana meningar, vilka var valda endast i hälften av resultaten. Således kan konstaterats att metoderna 1-8 åtminstone klarar av att filtrera bort meningar som inte ger något sakligt innehåll.

7.3. Prestanda

Skillnaden i prestandan för metoderna var ganska liten, vilket framgår ur följande tabell 7.3.2. Då prestandan mätts har inte tiden för själva förbehandlingen eller IDF uträkningen beaktats. Prestandan *P* är uträknad enligt följande:

$$P_m = \frac{\max(t)}{t_m} \quad , \qquad \textit{Formel 7.3.1.}$$

där t_m är tiden för metod *m,* och max(*t*) tiden för den långsammaste metoden.

Metod 1	Prestanda: 1.00064.23
Metod 2	Prestanda: 1.364
Metod 3	Prestanda: 1.011
Metod 4	Prestanda: 1.386
Metod 5	Prestanda: 1.030
Metod 6	Prestanda: 1.393
Metod 7	Prestanda: 1.032
Metod 8	Prestanda: 1.438

Tabell 7.3.2

Ur tabell 7.3.2 kan avläsas att metoderna 8, 6 och 4 är de snabbaste metoderna prestandamässigt. De är alltså de metoder som har minst antal tecken att jämföra mellan, då alla ”stop-words” är eliminerade, samt orden reducerade till sin stamform i metoderna 8 och 4.

8. Sammandrag och diskussion

8.1 Sammandrag

8 experiment utfördes för att 1: ta reda på om metoderna producerar vettiga resultat, och 2: för att jämföra hur mycket olika förbehandlingsmetoder av texten påverkar sammandraget.

Som konstaterats i 7.2 innehåller sammandragen från metoderna 1, 3, 5, 7 meningar som bidrar med någon form av substans ur innehållet, vilket inte alltid gällde för på måfå plockade meningar. Som också tidigare diskuterats är att resultaten i dessa experiment är svåra att jämföra med varandra och evaluera. Ur meningarna som extraherats i metoderna 1-8 kan tänkas att metoderna 1, 3, 5 och 7 ger de bästa resultaten, eftersom de inte innehåller meningar som är svåra att förstå på egen hand, men det är svårt att visa. Det räcker med att någon annan tycker olika, vilket många säkert gör, för att det skall bli nästan omöjligt att visa vem som har rätt.

Vad som däremot kan konstateras ur tabell 7.3.2., är precis som Porter föreslår [7], att det lönar sig prestandamässigt att reducera orden till sina stammar för att reducera komplexiteten då man bearbetar text.

Det som metoderna 2, 4, 6, 8 har gemensamt är att de eliminerar stop-words. Det här metoderna är även de metoder som enligt jämförelsen i 7.2 ger det sämsta resultatet för texten som sammanfattas.

8.2 Diskussion

Dessa statistiska metoder för att skapa sammandrag av text kräver nästan alltid att användaren redan är expert på textens innehåll för att kunna få något ut av sammandraget. De kan alltså oftast inte användas för att komprimera en längre text och ändå bevara de viktigaste tankarna. Olika typer av text bör även behandlas på olika sätt och med olika parametrar för att få bättre resultat. Dessa metoder och parametrar som använts bör dessutom kalibreras för hand för olika sorters text, vilket är svårt att göra om man inte är expert inom området.

Ett bättre område för användning av dessa metoder är alltså då användaren redan vet ungefär vad innehållet handlar om, t.ex. vid sökningar på sökmotorer, där sökmotorn visar en kort textsnutt som statistiskt bäst representerar innehållet av de sökord som angivits.

För att använda dessa metoder för att komprimera en längre text, bör det även kunna bevisas att texten innehåller de viktigaste idéerna. Detta område är ännu öppet för forskning.

Litteraturhänvisningar

1. René Arnulfo, García-Hernández, Romyna Montiel, Yulia Ledeneva, Eréndira Rendón, Alexander Gelbukh, RafaelCruz: Text Summarization by Sentence Extraction Using Unsupervised Learning. A.Gelbukh and E.F Morales (Eds.): MICAI 2008, LNAI5318, pp.133-143, Springer-Verlag Berlin Heidelberg 2008

2. Rada Mihalcea: Graph-based Ranking Algorithms for Sentence Extraction, Applied to Text Summaization, University of North Texas

3. Kiss, Tibor and Strunk, Jan (2006): Unsupervised Multilingual Sentence Boundary Detection. Computational Linguistics 32: 485-525.

4. Johnny Lindroos: Automatic Text Summarization Using MEAD, Experience with IMF Staff Reports, Turku 2006, PhD Thesis, Åbo Akademi.

5. Ramiz M. Aliguliyev: A new sentence similarity measure and sentence based extractive technique for automatic text summarization. R.M. Aliguliyev / Expert Systems with Applications 36 (2009) 7764–7772

6. Dragomir R. Radev, Hongyan Jing, Małgorzata Stys, Daniel Tam: Centroid-based summarization of multiple documents: Information Processing and Management 40 (2004) 919–938

7. M.F.Porter, An algorithm for suffix stripping; Originally published in Program, 14 no. 3, pp 130-137, July 1980

8. Satanjeev Banerjee, Ted Pedersen: The Design, Implementation and Use of the Ngram Statistics Package

9. Jade Goldstein, Jamie Carbonell; The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries

10. Källkoden ur MEAD v3.12: SimRoutines.pm daterat 11 Juni 2009; http://www.summarization.com/mead/