Uppsala universitet
Språkteknologiprogrammet
Diskurs, VT-99
Eva Forsbom
Lärare: Kari Fraurud

Automatisk sammanfattning av textinnehåll

Referat och studie av Gunnel Källgrens modell

BAKGRUND

Uppgiften - kort presentation

Rapporten handlar om en uppgift i kursen Diskurs på språkteknologiprogrammet. Uppgiften gick ut på att läsa en vetenskaplig artikel om en teori, modell eller metod för diskursanalys eller datoriserad diskursbearbetning, och att sedan referera den. Dessutom skulle man göra en egen empirisk studie över detta, antingen som en tillämpning, test eller utvärdering.

Jag valde att studera en artikel av Gunnel Källgren som handlar om automatisk generering av sammanfattningar, och att sedan testa modellen på en längre text.

Modellen - kort presentation

Modellen beskriver ett sätt att automatiskt generera en sammanfattning av en text med hjälp av innehållet i själva texten. Genom ord-, sats- och textanalys av källtexten kan man få ut så mycket information om innehållet i texten att det blir en grund för generering av en sammanfattning. Orden analyseras efter innehållsvikt, frekvens och samförekomst. Under satsanalysen klargörs relationerna mellan orden, och i textanalysfasen tas den övergripande textstrukturen fram. Utifrån den informationen genereras sedan sammanfattningen enligt önskat textmönster.

Studien - kort presentation

I min studie valde jag att generera en sammanfattning av Källgrens artikel, enligt den beskrivna modellen. Dels för att jag ville se hur modellen höll för en längre text, dels för att jag ändå skulle göra en sammanfattning av artikeln och för att det var bra att ha en manuellt sammanställd sammanfattning att jämföra de (halv)automatiskt framställda med.

Jag koncentrerade mig på ordanalyssteget i modellen och experimenterade med olika tröskelvärden för de ordanalyser som byggde på frekvensberäkning.

MODELLEN - ETT REFERAT

Modellens mål

I sin artikel *Automatic Absttracting of Content in Text* beskriver Gunnel Källgren en möjlig framtida modell för automatisk generering av textsammanfattningar som drar nytta av tekniker från flera forskningsområden inom lingvistik och artificiell intelligens. Möjlig i den meningen att alla problem ännu inte är lösta, men verkar kunna få en acceptabel lösning inom en överskådlig framtid.

Målet med modellen är att genom ord-, sats- och textanalys av en källtext få så mycket information om innehållet i texten att det blir en grund för generering av en sammanfattning. Orden analyseras efter innehållsvikt, frekvens och samförekomst, och bara de meningar som innehåller de viktigaste orden går vidare till nästa steg, satsanalysen. Där klargörs relationerna mellan orden, och i textanalysfasen tas den övergripande textstrukturen fram. De olika analysnivåerna är beroende av varandra, så om det saknas information i ett steg bör det vara möjligt att gå tillbaka och samla in mer information, men ordanalysen är den inledande och styrande faktorn.

När dessa analyser är utförda ska det finnas tillräckligt med information att generera en sammanfattning av lämplig omfattning. Det kan ske genom återvinning av hela meningar från källtexten, genom komprimering av meningar från källtexten så att inga ”onödiga” ord finns kvar eller genom nya meningar utifrån de representationer av innehållet som är resultatet från de tidigare analyserna enligt önskat textmönster.

Modellens moduler

1 Ordanalys

1.1 Ord och innehåll

Orden kan liksom meningarna och texten ses som en integrerad del av textens innehåll, eftersom texten får sin betydelse utifrån de ord den består av, medan orden och meningarna får sin betydelse utifrån den kontext de står i. När man ska frekvensberäkna och välja ut de ord som ska representera innehållet i texten räcker det enligt Källgren för den automatiska sammanfattningsgeneratorn att bara räkna med ett ords grundbetydelse och skippa finliret i betydelseskillnader mellan olika kontexter, eftersom den tilltänkta läsaren själv kan göra den betydelsetolkningen utifrån kontexten.

Däremot är det viktigt att välja ut vilka sorters ord som ska representera textens innehåll. Vissa sorter bär t.ex. på mer innehåll än andra och bör finnas med, medan andra är relativt innehållstomma och kan tas upp i en stopplista och räknas bort vid frekvensberäkningen. Ord som ofta kan avfärdas är de inom de slutna ordklasserna, men inte alltid. Pronomen kan t.ex. referera till viktiga begrepp för texten och kan inte helt lättvindligt ignoreras, utan kanske istället borde ges extra vikt. De kan t.ex. bytas ut mot sina korrelat enligt en pronomenupplösningsalgoritm, så att begreppen får sin rätta vikt.[1] Dessutom anser Källgren att adverb som negationer ska tas med bland innehållsorden eftersom de ändrar innehållet i så stor betydelse.[2] Däremot kan vissa platshållarverb med lite innehåll med fördel förpassas till stopplistan, exempelvis *ha*, *vara*, *komma* och *få*.

[1] Jfr min anmärkning under Pronomen s. 7.
[2] Jfr min anmärkning under Stopplista s. 8.

1.2 Ord och frekvenser

I resultatet av frekvensberäkningen brukar man finna de ord som verkar bäst representera innehållet i en text i mellanfrekvensintervallet.

Ord med väldigt hög frekvens har ofta ett litet betydelseinnehåll. Det gäller exempelvis formorden, som filtreras bort via stopplistan, men det gäller även särskilda ord inom vissa texttyper eller -domäner, t.ex. *experiment* inom vetenskapliga texter. Ett sätt att filtrera bort även dem skulle, enligt Källgren, vara att skapa domänspecifika stopplistor som man sedan, enligt min tolkning, antingen aktiverar manuellt eller som aktiveras automatiskt av den signal som ges av att orden i dem förekommer väldigt ofta i texten.

Ord med väldigt låg frekvens tar upp en stor del av textens volym, men beskriver oftast inte det centrala innehållet i texten och kan därför oftast räknas bort. Men vissa ord som förekommer sällan kan tillhöra kärnvokabulären inom en viss texttyp eller -domän och borde alltid uppmärksammas. Man skulle kunna ta med dem i domänspecifika ta-med-listor på motsvarande sätt som för stopplistorna.

Hur många ord i mellanfrekvensintervallet som ska gå vidare till nästa fas i analysen beror t.ex. på hur lång källtexten är och hur lång sammanfattningen ska vara.

1.3 Ord och kollokationer

En annan intressant faktor att ta i beaktande är kollokationer, dvs. ord som förekommer tillsammans oftare än väntat, sett till deras frekvens i löpande text. I en storskalig studie har Martin Phillips visat att kollokationer av innehållsord som förekommer mycket oftare än väntat kan ge en ganska god beskrivning av innehållet i längre vetenskapliga texter. Med kollokationer menar han då ett innehållsord och alla innehållsord som förekommer inom fyra ord före eller efter det (utan avbrott av skiljetecken).

Vid analysen för han då alla besläktade ordformer till samma lemma, alltså även ordformer från olika ordklasser, t.ex. *representera*, *representativ* och *representant*. Det gör även Källgren i det manuella analysexempel i den senare delen av rapporten som ska illustrera modellen. Där ser man hur även vokabulärens interna struktur kan ge ledtrådar om textens huvudinnehåll.

2 Satsanalys

Enbart vokabulärens struktur kan dock inte ge en helhetsbild av textinnehållet, utan det måste även till sats- och textanalys. Satsanalysen är tänkt att utföras av en parser, vilket ställer till vissa problem i nuläget. Ingen av de parsrar som finns idag klarar nämligen att analysera texter som inte är begränsade till vare sig vokabulär eller domän, men modellen är tänkt att kunna skapa sammanfattningar även av sådana texter.

Ett annat problem är representationen av analysresultatet från parsern. Den kan se ut på flera olika sätt, men gemensamt för de olika sätten är att de har små möjligheter att visa relationer som sträcker sig över meningsgränser. Sådana relationer är ofta lika viktiga för textinnehållet som relationer inom meningar.

Eftersom modellens analys av innehållet ska ske i flera steg, på ord-, sats- och textnivå, kanske det inte är nödvändigt att parsa varje mening fullt ut, utan det kanske räcker med dem som ordanalysen pekar ut som centrala. Annars kan det räcka med en partiell parsning för att spåra de centrala ordens relationer i den övriga texten.[3]

Satsanalysen skulle också kunna filtrera bort sådana ord som fyller ut och kittar ihop en text, men som är överflödiga i en sammanfattning.

[3] Jfr min anmärkning under Pronomen s. 7.

3 Textanalys

Analysen på textnivå är tänkt att identifiera koreferensrelationer samt centrala teman i texten. Inte heller där har dock forskningen kommit så långt att en sådan är genomförbar för obegränsad naturlig text, huvudsakligen för att textanalys kräver så pass mycket arbete och att de resultat man ändå får fram kan vara svåra att generalisera.

En ljuspunkt för området automatisk sammanfattning gäller sambandskedjor, dvs. kedjor av ord som refererar till samma eller närbesläktade begrepp. De spelar en viktig roll för strukturering av texten, och begreppens ”förmåga” att ingå i sådana kedjor speglar ofta hur mycket de bidrar till textens semantiska innehåll.

Dessa sambandskedjor skulle kunna byggas upp med hjälp av något slags tesaurus, med synonymer, antonymer samt hypo- och hyperonymer. Liksom vid ordanalysen går det att lemmatisera över ordklassgränserna. Dessutom skulle det vara möjligt att göra ett slags semantisk lemmatisering utifrån de semantiska relationerna. Källgren hävdar att det inte skulle vara fråga om några världskunskapsdatabaser, utan att den semantiska analysen ska vara kopplad till ordanalysen.[4]

Även textens grafiska uppbyggnad ger ledtrådar om innehållet. Det kan t.ex. gälla indelning i kapitel och stycken. Innehållsord som förekommer i rubriker beskriver ofta textinnehållet mer än andra.

4 Textgenerering

När analyserna är klara ska det finnas kvar ett skelett av källtexten med de viktigaste meningarna som man sedan ska utgå ifrån vid genereringen av sammanfattningen.

Ett slags sammanfattning skulle kunna vara att helt enkelt visa en lista med alla innehållsord eller en strukturerad lista med alla kollokationer. Sådana sammanfattningar kan vara användbara vid automatisk indexering, och de kan ge en vink om relationer i texten, men är ganska svåröverblickbara.

Syftet med den här modellen är dock att använda det framanalyserade skelettet till att i en särskild textgenereringskomponent skapa en sammanfattning uttryckt i naturligt språk. Om relationerna i skelettet är lagrade på ett lämpligt sätt i en databas så att det går att hämta dem och uttrycka dem i ord kunde man kanske använda tekniker från expertsystem och fråga-svardatabaser, där man genererar sammanhängande text utifrån fakta i databasen. Men det finns vissa risker man bör beakta i sådana fall: det kan saknas information som gör att sammanfattningen blir helt missvisande, t.ex. om negerande ord från texten saknas, och fråga-svarsystem har oftast ingen längdbegränsning på svar utan de blir så långa att alla fakta som ska komma med kommer med. Sammanfattningarnas längd bör dock gå att variera efter källtextens längd och innehållstäthet.

I det exempel Källgren använder utnyttjar hon en annan metod, nämligen att lagra de centrala meningarna i den form de är uttryckta i källtexten. Vid genereringen återanvänds de som de är eller med vissa förändringar. Förändringarna kan t.ex. vara att ”pratiga” och redundanta meningar rensas från ”pratet” och redundansen samt att relationerna mellan meningarna tydliggörs så att meningarna kittas samman till en text.

[4] Men även om man utgår från texten och representerar den textinterna semantiska strukturen måste väl kunskapen vara representerad någonstans så att det går att söka i den för alla textsorter?

Modellen - ett exempel

Sist i artikeln ger Källgren ett konkret exempel på hur en sammanfattning skulle kunna genereras genom att för hand analysera en kort vetenskaplig text på 246 ord enligt modellen. Koncentrationen ligger på ordanalysdelen, eftersom den är mest genomförbar i dagsläget.

Först tar hon ut innehållsorden, och byter ut pronomen mot motsvarande nominalfras (visas i texten som versaler). Det resulterar i en lista över innehållsord på 128 ord.

Därefter samlar hon orden till lemman (enligt beskrivningen i modellen) och tar ut kollokationer. Bara kollokationer som förekommer fler än en gång kommer med (om de bara förekommer en gång är de ju slumpmässiga). Hade det varit en längre text kunde frekvensgränsen ha varit högre. Eftersom ett innehållsord då måste förekomma i två kollokationer räcker det att ta ut kollokationer för ord med frekvensen två eller högre. Version 2 av sammanfattningen består av alla meningar som innehåller en kollokation som har frekvensen två eller högre i texten. Versionen innehåller 107 ord.

Sedan tar hon bort allt onödigt prat och får en version på 86 ord, innan hon ”tolkar” och ”förstår” den förkortade texten och får slutversionen, som består av 27 ord.

I exemplet visar Källgren att det i stort skulle gå att få fram version 2 på automatisk väg i dag, och att all information som behövs för de senare versionerna, där bl.a. en parser och semantisk komponent behövs, finns i den versionen, så de mer resurskrävande analyserna kan utgå därifrån.

(Sammanfattningarna visas i Bilaga 1 - Modellexemplet i olika versioner.)

MODELLEN - EN STUDIE

Studiens mål

I denna uppgift ingick att utöver sammanfattningen av modellen även göra någon form av implementering av (åtminstone någon del av) den. Jag valde att liksom Källgren göra en analys för hand (i stort sett) av en vetenskaplig text, men en betydligt längre text (7 476 vs 247 ord), nämligen Källgrens artikel. Det gjorde jag av tre skäl:

- Jag ville se vilka problem som dök upp vid en så pass mycket längre text. Texttypen behöll jag konstant för att inte blanda ihop texttypens och textlängdens påverkan på resultatet av jämförelsen.
- Jag skulle ändå göra en sammanfattning av texten, så det finns en manuellt sammanställd sådan att jämföra med för den som så önskar.
- Jag hade inte tillgång till de verktyg som behövdes för en automatisk analys, inte ens de som enligt Källgren finns i dagsläget. Men genom en närkamp med text och modell hoppades jag komma på idéer om hur analysen kunde automatiseras.

Liksom Källgren koncentrerade jag mig på ordanalyssteget i modellen eftersom de andra analyserna än så länge inte är genomförbara för längre text i nuläget.

Ordanalys

Även om analysen mestadels skedde för hand kunde jag ändå utnyttja datorkraft till de mer mekaniska delarna av analysen, och inledde därför arbetet med att göra texten maskinläsbar[5]. Eftersom jag skulle använda den elektroniska versionen i ett textbehandlingsprogram gjorde jag ingen skillnad på rubriker och annan text. En sådan skillnad kunde annars ha utnyttjats som information om innehållsvikt eftersom rubriker ofta bär mycket av det centrala innehållet i texter. Det nämner även Källgren i artikeln.

1 Ord och innehåll

Pronomen

För ordanalysen behövde jag segmentera texten till enskilda ord, och för det använde jag programmet TSSA[6]. Men dessförinnan bytte jag ut anafora pronomen mot deras fullformer (visas med versaler i texten). Jag följde inte (inte medvetet, i varje fall) den algoritm för pronomenupplösning[7] som Källgren nämnde i artikeln eftersom den var utformad för svenska och artikeln var på engelska (vilket exempeltexten i artikeln också var, så jag tror inte att Källgren använde algoritmen heller).

Det största problemet var inte heller de personliga pronomenen som hade nominalfraser som antecedenter, utan de pronomen som syftade på satser, och de hade algoritmen ändå inte kunnat lösa. Men om man inte löser upp dem och den sats de syftar på försvinner vid reduceringen av texten blir de ju helt missvisande, så det problemet måste få en lösning om modellen ska kunna fungera. Förutom pronomen finns det även andra ord som syftar framåt och bakåt i texten och måste tas om hand på liknande sätt, t.ex. *former* och *latter*, som jag inte heller tänkte på att byta ut[8].

[5] Jag skrev in den för hand, vilket resulterade i ett antal skrivfel som jag försökte upptäcka och rätta till, men några blev tyvärr kvar. De är dock av minimal betydelse för analysen.

[6] Se Dahlqvist 1994. Jag använde standarduppsättningen med ordavgränsare: punkt, kommatecken, kolon, semikolon, frågetecken, utropstecken, citattecken, vänster- och högerparentes samt snedstreck.

[7] Se Fraurud 1988.

[8] Dessutom fuskade jag lite till, och brydde mig inte om relativa eller reflexiva pronomen. De borde naturligtvis också bytas ut, i konsekvensens namn, så att frekvensen för referenterna blir riktig. Däremot är det mindre risk för missförstånd för dem eftersom deras referenter oftast finns i samma mening.

Skulle en sådan pronomenupplösning göras automatiskt för de personliga pronomenen skulle det behövas mer än ytlig parsning, och för upplösning av satssyftande pronomen och liknande uttryck skulle det behövas än rejälare doningar. Här anar jag ett cirkelresonemang i artikeln. Pronomenupplösningen skulle enligt modellen ske före frekvensberäkningen så att referenterna fick den frekvens de skulle ha, och den noggrannare parsningen skulle ske först sedan texten reducerats baserat på frekvensberäkningen så att det inte krävdes så mycket resurs- och kunskapskrävande parsning. Men eftersom det krävs noggrann parsning för pronomenupplösningen görs aldrig den resursbesparingen. Frågan är dock om pronomenupplösningen behöver göras före frekvensberäkningen. För längre texter än exempeltexten i artikeln borde de centrala innehållsorden förekomma tillräckligt ofta ändå för att tas med i den vidare beräkningen. Man skulle ju kunna lösa upp pronomenen först i satsanalyssteget eftersom all information ändå ska finnas kvar så att det går att söka i den. Det kanske kräver lite mindre resurser i varje fall.

Stopplista

När pronomenupplösningen och segmenteringen i ord var avklarad tog jag för hand[9] ut den stopplista över ord som inte skulle tas med i frekvensberäkningen, dvs. formord, innehållslätta platshållarverb och liknande. Till skillnad från Källgren tycker jag att satsadverb som *not* och liknande gott kan vara med i stopplistan. Man behöver inte ta med dem vid frekvensberäkningen och i kollokationerna, utan det räcker med att de kommer med i satsanalyssteget eftersom de bestämmer hela satser och det är hela satser som tas ut för analys (åtminstone i den variant av modellen som användes för exemplet) inte enskilda ord. Dessutom ska ju all information ändå ska finnas kvar så att det går att söka i den.

Stopplistan visas i Bilaga 2 - Stoppord.

2 Ord och frekvenser

Jag använde ingen särskild domän- eller texttypsspecifik stopplista, men tog med några innehållslösa perceptionsverb, t.ex. *see* och *seem*, i den redan nämnda listan. Andra ord som kunde ha ingått i en texttypsstopplista är exempelvis *represent*, *interesting*, *method*, *model*, *result*, *relevant* och *significant*, dvs. ord som handlar om själva undersökningsmetoden, om teorimodeller och värdering av dessa. I en domänstopplista hade jag kunnat ha med t.ex. *text*, *language* och *linguistic*. *Text* var t.ex. det innehållsord som förekom oftast i texten (71 gånger) och resultatet hade nog blivit ganska så annorlunda om jag använt sådana stopplistor, men en sådan undersökning får anstå till ett annat tillfälle.[10]

När jag skulle lemmatisera orden utgick jag från en alfabetiskt sorterad frekvenslista över innehållsorden som jag skapat i TSSA. Den gjorde ingen åtskillnad mellan homografer, och det gjorde inte jag heller. Om man ska skilja mellan homografer automatiskt måste man på samma sätt som för pronomenupplösningen parsa texten innan frekvensberäkningen, men den skillnaden att för homografupplösning behövs bara ytlig parsning. Kanske kan man också spara homografupplösningen till satsanalyssteget och då ta bort satser som innehåller en ointressant homograf till ett intressant innehållsord.

Jag arbetade främst med initial stränglikhet när jag lemmatiserade över ordgränserna, men ibland arbetade jag mer intuitivt, som vid *mid-frequency och organic-looking*, och ibland kunde jag inte bestämma mig utan räknade ordet två gånger, som vid *abstract-*

[9] Dvs. jag gjorde bedömningen av vilka ord som skulle vara med i listan efter eget huvud. Däremot utgick jag från en initialsorterad frekvenslista över ordförekomsterna i texten som jag skapat i TSSA. Stopplistan använde jag sedan som TSSA-stopplista mot filen med de segmenterade orden för att få fram en lista över innehållsorden.

[10] Källgren nämner i artikeln att innehållsorden i texttypsspecifika stopplistor är så vanligt förekommande i texttypen att de nästan blir innehållslösa, utom som signal om texttypen i fråga. Då kom jag osökt att tänka på att man istället för att använda orden i en stopplista först kunde utnyttja den 'signalen' från frekvensberäkningen och sända den till textgenereringskomponenten som val av texttyp. Sedan kunde man ta bort orden från frekvenslistan. Det skulle för övrigt vara intressant att jämföra olika texttyper och se vilka ord som var texttypiska och hur ofta dessa ord måste förekomma i texten för att tappa så mycket i innehåll att de blir texttypiska.

generating och *question-answering*. Så går det naturligtvis inte att göra i ett automatiskt system, utan då måste man bestämma sig för en metod. Kanske initial stränglikhet är bäst, åtminstone för engelska och svenska. Man missar visserligen några förekomster, men om lemmat är tillräckligt intressant kommer det nog med till finalen ändå.

För den korta exempeltexten i artikeln valde Källgren att sätta tröskelvärdet till 2. Min text var betydligt längre, så valde jag ett högre värde, till att börja med 5. Eftersom jag var intresserad av vad ett lämpligt värde kunde vara provade jag sedan med flera olika värden, men mer om det senare.

I Bilaga 3 - Ord med frekvens ≥ 5 (alfabetisk ordning) visas min sortering i lemman i alfabetisk ordning och även vilka ord jag fört till de lemmana. I Bilaga 4 - Ord med frekvens ≥ 5 (frekvensordning) visas enbart lemmana i frekvensordning.

3 Ord och kollokationer

För de ord som hade en frekvens över 5 kontrollerade jag sedan om de förekom tillsammans med något av de andra orden med frekvens över 5 inom 4 ord. Tyvärr upptäckte jag sedan jag gjort analysen att jag tolkat intervallet fel och att man inte skulle räkna med det ord man utgick ifrån i spannet utan räkna ordet plus 4 ord till[11]. Det påverkar naturligtvis resultatet på så sätt att det saknas några kollokationer, men de jag fick fram skulle ändå ha kommit med.

Jag provade två metoder. Först gjorde jag en 4-gramlista av texten i TSSA och sökte och ersatte alla ord som inte fanns med i frekvenslistan eller hade en frekvens under 5 med @. Problemet med den metoden var att vissa förekomster räknades flera gånger, och jag kom inte på något bra sätt att komma runt det. Då skapade jag en konkordans för orden i frekvenslistan i TSSA och tog ut tre ord efter varje ord. Sedan sökte jag och ersatte med @ på samma sätt som ovan. På så sätt räknades förekomsterna bara en gång ivarjefall. (I ett mer integrerat system skulle man naturligtvis ha märkt orden med etiketter för lemman och lagrat dem och frekvensinformationen i en databas, där man sedan kunde söka efter samförekomsterna och i sin tur märka dem med egna etiketter.)

Överhuvudtaget uppstod flera frågetecken kring kollokationerna, som det inte fanns något svar på i artikeln. Hur skulle jag t.ex. göra med två kollokationer där samma ord förekom i samma ordning men på olika positioner, som *PARSE @ SENTENCE @* och *PARSE SENTENCE @ @*? Skulle jag räkna dem som en eller två kollokationer? Och hur skulle jag göra med kollokationer med fler än två innehållsord, som *AUTOMATIC ABSTRACT @ CONTENT*, där två av orden även förekom tillsammans minst en gång utan det eller de andra orden? Skulle jag räkna den som en treordskollokation eller varje tvåordskollokation för sig och på så sätt öka frekvensen för dem? Jag valde det senare alternativet för båda. I efterhand ser jag att det hade varit bättre att välja det första alternativet för den första frågan eftersom orden ivarjefall förekommer inom samma intervall.

I Bilaga 5 - Kollokationer för Ord med frekvens ≥ 2 visas de kollokationer jag fick fram och den frekvens de hade i texten.

[11] Jfr Källgren 1990.

Sammanfattningarna

När ordanalysen var klar hade jag den information jag behövde för att kunna ta ut de meningar som verkade vara centrala i texten. Jag gjorde först en meningsfil i TSSA, och använde standarduppsättningen för meningsavgränsare. Därefter märkte jag kollokationerna med etiketter, och tog ut de meningar som innehöll kollokationer med en viss frekvens.

Jag provade sedan med olika tröskelvärden för ord- och kollokationsfrekvenserna. Antalet ord i de resulterande sammandragen visas i tabellen nedan, och tre av sammandragen visas i bilagorna 6-8. Sammandraget i bilaga 6 är för långt, nästan hälften är kvar, och innehåller fortfarande överflödig information. Sammandraget i bilaga 8 är tillfredsställande kort, men saknar en del viktig information som man isåfall måste hämta fram ur texten igen, beroende på hur detaljerad sammanfattning man vill ha. Sammandraget i bilaga 7 är fortfarande för långt, men har i stort sett all information som behövs för att bygga en sammanfattning ifrån. Jag använde t.ex. den som grund för min sammanfattning, som är 1 681 ord lång.

Ordfrekvens	Kollokationsfrekvens	Antal ord i sammandrag
1	1	7 476[12]
5	2	4 998
5	3	4 129
5	4	3 633
5	5	3 059[13]
6	2	4 831
7	2	4 831
8	2	4 617
9	2	4 093
10	2	3 937
11	2	3 418
12	2	3 366
14	2	2 922
15	2	2 453
16	2	2 308
17	2	1 851
18	2	1 190
19	2	1 045
20	2	991[14]
10	3	3 420
10	4	3 131
10	5	2 549
10	6	2 136[15]

[12] Källtexten.
[13] Se Bilaga 6 - Sammanfattning 1 (3 059 ord).
[14] Se Bilaga 8 - Sammanfattning 3 (991 ord).
[15] Se Bilaga 7 - Sammanfattning 2 (2 136 ord).

REFERENSER

Dahlqvist, Bengt. 1994. *TSSA 2.0*. A PC Program for Text Segmentation and Sorting. Institutionen för lingvistik, Uppsala universitet.

Fraurud, Kari. 1988. Pronoun Resolution in Unrestricted Text. I: *Nordic Journal of Linguistics*, Vol. 11(1-2): 47-68.

Källgren, Gunnel. 1988. Automatic Abstracting of Content in Text. I: *Nordic Journal of Linguistics*, Vol. 11(1-2): 89-110.

Källgren, Gunnel. 1990. Automatic Indexing and Generating of Content Graphs from Unrestricted Text. I*: Papers from the Seventh Scandinavian Conference of Computational Linguistics*. Reykavik. 60-76.

BILAGA 1 - MODELLEXEMPLET I OLIKA VERSIONER

Originalversion

With the aid of a relatively simple program a computer can be converted into a kind of microscope for viewing the boundary of the Mandelbrot set. In principle one can zoom in for a closer look at any part of the set at any magnification. From a distant vantage the set resembles a squat, wart-covered figure eight lying on its side. The inside of the figure is ominously black. Surrounding it is a halo-colored electric white, which gives way to deep blues and blacks in the outer reaches of the plane.

Approaching the Mandelbrot set, one finds that each wart is a tiny figure shaped much like the parent set. Zooming in for a close look at one of the tiny figures, however, opens up an entirely different pattern: a riot of organic-looking tendrils and curlicues sweeps out in whorls and rows. Magnifying a curlicue reveals yet another scene: it is made up of pairs of whorls joined by bridges of filigree. A magnified bridge turns out to have two curlicues sprouting from its center. In the center of this center, so to speak, is a four-way bridge with four more curlicues, and in the center of these curlicues another version of the Mandelbrot set is found.

The magnified version is not quite the same Mandelbrot set. As the zoom continues, such objects seem to reappear, but a closer look always turns up differences. Things go on this way forever, infinitely various and frighteningly lovely.

Version 1

aid relatively simple program computer converted microscope viewing boundary Mandelbrot set. principle zoom closer look part set magnification. distant vantage set resembles squat, wart-covered figure eight lying FIGURE side. inside figure ominously black. Surrounding FIGURE halo-colored electric white, HALO way deep blues blacks outer reaches plane.

Approaching Mandelbrot set, finds wart tiny figure shaped much like parent set. Zooming close look tiny figures, opens entirely different pattern: riot organic-looking tendrils curlicues sweeps whorls rows. Magnifying curlicue reveals scene: made pairs whorls joined bridges filigree. magnified bridge turns have curlicues sprouting BRIDGE center. center center, four-way bridge more curlicues, center curlicues version Mandelbrot set found.

magnified version not quite Mandelbrot set. zoom continues, objects seem reappear, closer look always turns differences. Things way forever, infinitely various frighteningly lovely.

Version 2

In principle one can zoom in for a closer look at any part of the MANDELBROT SET at any magnification. From a distant vantage the set resembles a squat, wart-covered figure of eight. Approaching the Mandelbrot set, one finds that each wart is a tiny figure shaped much like the parent set. Zooming in for a closer look at one of the tiny figures, however, opens up an entirely different pattern. A magnified bridge turns out to have two curlicues sprouting from its center. In the center of these curlicues another version of the Mandelbrot set is found. The magnified version is not quite the same Mandelbrot set.

Version 3

One can zoom in for a closer look at any part of the MANDELBROT SET at any magnification. From a distant vantage the set resembles a squat, wart-covered figure. Each wart is a tiny figure shaped much like the parent set. Zooming in for a close look at one of the tiny figures opens up a different pattern. A magnified bridge has two curlicues from its center. In the center of these curlicues another version of the Mandelbrot set is found. The magnified version is not the same Mandelbrot set.

Slutversion

Any part of the Mandelbrot set can be magnified. The magnified version will be a figure much like the first Mandelbrot set but not quite the same.

BILAGA 2 - STOPPORD

08
1
10-12
11
14
19
1936
1963
1970
1972
1974
1976
1977
1978
1979
1982
1983
1984
1985
1988
1a
1b
2
21
246
3
39
4
43
5
6a
6b
7
8
89-110
91
a
about
above
according
actually
after
ago
al
all
along
also
always
among
an
and
another
any
are
as
at
august
away
b
be
become
been
before
being
below
beside
between
both
but
by
c
can
cannot
certain
cf
come
comes
could
do
does
done
down
due
e
each
ed
eight
els
enough
entirely
et
even
every
except
fairly
felt
few
find
'find'
finding
finds
first
following
for
forever
found
four
from
full
fully
further
g
get
getting
give
given
gives
giving
go
h
had
has
have
here
highly
how
however
i
if
immediat
e
immediat
ely
in
inside
into
is
it
itself
j
just
k
least
less
let
made
many
may
might
more
most
mostly
much
must
my
n
next
no
none
not
now
of
often
on
one
only
or
other
others
our
out
outer
over
partly
perhaps
probably
quite
rather
really
s-106
said
same
see
seem
seemed
seems
seen
septembe
r
several
should
shown
so
so-called
some
somethin
g
sometime
s
speak
still
such
t
take
taken
taking
than
that
the
then
there
these
this
those
though
three
thus
to
together
too
towards
tried
two
under
up
us
use
used
using
various
we
very
what
when
where
whether

which	with	would	y
while	within	x	yet
will	without	x2	you

BILAGA 3 - ORD MED FREKVENS ≥ 5 (ALFABETISK ORDNING)

18 abstract x 4
- abstracting (3)
- abstraction (2)
- abstracts (7)
- abstract's (1)
- abstract-generating (1[1])

19 analysis (16)
- analysed (1)
- analysing (1)
- analysis' (1)

7 answer (1)
- answers (3)
- question-answering (3[2])

5 appear (2)
- appearance (1)
- appearing (2)

6 article (5)
- articles (1)

14 automatic (9)
- automatically (1)
- automatized (4)

5 central (5)

14 close (6)
- closer (7)
- 'close (1)

17 collocate (1)
- collocates (1)
- collocation (6)
- collocational (1)
- collocations (8)

25 compute
- computational (5)
- computer (15)
- computerized (4)
- computer-based (1)

5 connect (1)
- connection (4)

6 construct (1)
- constructed (1)
- constructing (3)
- construction (1)

27 content (24)
- contents (2)
- 'content (1)

6 curlicue (4)
- curlicues (2)

5 develop
- developed (1)
- developing (1)
- development (3)

13 differ
- differences (3)
- different (9)
- differs (1)

5 distant (5)

14 figure (12)
- figures (2)

8 form (6)
- forms (2)

17 frequency (8)
- frequencies (4)
- frequent (3)
- mid-frequency (2)

17 generate (1)
- generated (1)
- generating (4)
- generation (7)
- generator (3)
- abstract-generating (1[3])

9 good (8)
- better (1)

5 grammar (1)
- grammatical (4)

5 gunnel (5)

5 idea (5)

10 importance (3)
- important (7)

6 information (6)

7 interest (2)
- interesting (5)

7 kind (4)
- kinds (3)

6 källgren (6)

13 language (12)
- languages (1)

6 lemma
- lemmas (4)
- lemmatization (1)
- lemmatized (1)

9 like (9)

13 linguistic (1)
- linguistics (12)

7 list (2)
- listed (1)
- lists (4)

8 long (4)
- longer (4)

19 look (15)
- look' (1)
- organic-looking (3)

15 magnify (3)
- magnification (5)
- magnified (5)
- magnifying (2)

16 mandelbrot (14)
- 'mandelbrot (1)
- 'mandelbrot' (1)

14 mean
- meaning (7)

[1] Räknad två gånger.

[2] Räknad två gånger.

[3] Räknad två gånger.

meanings (1)
means (5)
meant (1)
8 model (7)
model's (1)
14 natural (14)
6 necessary (6)
5 need (2)
needed (3)
9 occur (2)
occurrence (2)
occurrences (3)
occurs (1)
co-occur (1)
7 output (7)
26 parse (2)
parsed (2)
parser (4)
parsers (3)
parsers' (1)
parsing (13)
'parsing' (1)
16 part (14)
parts (2)
5 pattern (3)
patterns (2)
5 phillips (4)
phillips' (1)
6 principle (5)
principles (1)
8 process (2)
processes (3)
processing (3)
5 program (4)
program's (1)
5 question (1)
questions (1)
question-answering (3[4])
5 read (2)
readability (1)
readable (2)
5 reduce
reduced (4)
reduction (1)
8 regard (2)
regarded (4)
regardless (2)
9 represent (1)
representation (1)
representatives (2)
represented (1)
representing (4)
5 science (1)
scientific (4)
8 semantic (4)
semantically (2)
semantics (2)
15 sentence (3)
sentences (10)
sentences' (1)
sentential (1)
25 set (23)
sets (1)
set' (1)
7 short (5)
shorter (1)
shortness (1)
7 simple (4)
simplest (1)
simply (2)
10 source (9)
sources (1)
5 step (4)
steps (1)
5 stockholm (4)
stockholms (1)
8 structure (5)
structuring (2)
structural (1)
20 system (10)
systems (8)
systems' (2)
71 text (52)
texts (12)
textual (4)
text's (2)
non-textual (1)
9 tiny (9)
11 wart (7)
wart-covered (4)
22 way (17)
ways (5)
5 verbs (5)
11 version (11)
10 vocabulary (10)
62 word (13)
words (46)
words' (2)
word' (1)
14 zoom (9)
zooming (4)
'zoom' (1)

[4] Räknad två gånger.

BILAGA 4 - ORD MED FREKVENS ≥ 5 (FREKVENSORDNING)

71 text

62 word

27 content

26 parse

25 compute
25 set

22 way

20 system

19 analysis
19 look

18 abstract

17 collocate
17 frequency
17 generate

16 mandelbrot
16 part

15 magnify
15 sentence

14 automatic
14 close
14 figure
14 mean
14 natural
14 zoom

13 differ
13 language
13 linguistic

11 wart
11 version

10 importance
10 source
10 vocabulary

9 good
9 like
9 occur
9 represent
9 tiny

8 form
8 long
8 model
8 process
8 regard
8 semantic
8 structure

7 answer
7 interest
7 kind
7 list
7 output
7 short
7 simple

6 article
6 construct
6 curlicue
6 information
6 källgren
6 lemma
6 necessary
6 principle

5 appear
5 central
5 connect
5 develop
5 distant
5 grammar
5 gunnel
5 idea
5 need
5 pattern
5 phillips
5 program
5 question
5 read
5 reduce
5 science
5 step
5 stockholm
5 verbs

BILAGA 5 - KOLLOKATIONER FÖR ORD MED FREKVENS ≥ 2

12 CLOSE LOOK @ @

9 CONTENT @ @ TEXT
9 MANDELBROT SET @ @

8 NATURAL LANGUAGE @ @

7 CONTENT WORD @ @
7 FORM WORD @ @
7 LIST @ @ WORD
7 SOURCE TEXT @ @
7 WORD @ @ TEXT

6 GENERATE @ @ TEXT
6 TINY FIGURE @ @
6 VOCABULARY ANALYSIS @ @

5 ABSTRACT @ CONTENT @
5 GENERATE @ ABSTRACT @
5 MAGNIFY VERSION @ @
5 REDUCE VERSION @ @
5 TEXT LINGUISTIC @ @

4 AUTOMATIC ABSTRACT @ CONTENT
4 COLLOCATE @ @ WORD
4 COMPUTE LINGUISTIC @ @
4 CONTENT @ TEXT @
4 FIGURE @ @ LIKE
4 FREQUENCY WORD @ @
4 GENERATE ABSTRACT @ @
4 LANGUAGE TEXT @ @
4 SET @ @ MAGNIFY
4 VERSION @ @ MANDELBROT
4 WORD @ @ WORD
4 WORD FREQUENCY @ @
4 WORD OCCUR @ @

3 ANSWER SYSTEM @ @
3 COMPUTE GENERATE @ @
3 CURLICUE @ VERSION @
3 DEVELOP @ COMPUTE LINGUISTIC
3 DIFFER PART @ @
3 DISTANT @ @ SET
3 KIND @ ANSWER @
3 LIKE @ @ SET
3 LOOK @ @ PART
3 MEAN @ @ WORD
3 MODEL @ @ GENERATE
3 PART @ @ MANDELBROT
3 PART @ @ TEXT
3 QUESTION SYSTEM @ @
3 SENTENCE @ CENTRAL @
3 TEXT ANALYSIS @ @
3 WART @ @ TINY
3 WART FIGURE @ @
3 WAY @ REPRESENT @
3 WORD APPEAR @ @

2 ABSTRACT @ @ READ

2 ABSTRACT @ TEXT @
2 ANALYSIS @ @ VOCABULARY
2 ANALYSIS @ NATURAL TEXT
2 AUTOMATIC @ GENERATE @
2 AUTOMATIC GENERATE @ @
2 AUTOMATIC MEAN @ @
2 CENTRAL @ NEED @
2 CENTRAL CONTENT @ @
2 COLLOCATE @ WORD @
2 COLLOCATE APPEAR @ @
2 COLLOCATE FREQUENCY @ @
2 DEVELOP FORM @ @
2 DIFFER @ @ SEMANTIC
2 DIFFER @ @ ANALYSIS
2 FIGURE @ @ FIGURE
2 FORM @ CONTENT @
2 FREQUENCY @ @ TEXT
2 FREQUENCY @ @ WORD
2 FREQUENCY @ WORD @
2 GENERATE @ @ ABSTRACT
2 GENERATE @ @ SENTENCE
2 GENERATE @ NATURAL @
2 GRAMMAR FORM WORD @
2 IMPORTANCE @ @ WORD
2 INFORMATION @ @ SOURCE
2 KIND @ TEXT @
2 LIST @ COLLOCATE @
2 LONG TEXT @ @
2 MEAN @ @ TEXT
2 NEED @ @ ANALYSIS
2 OCCUR @ @ TEXT
2 OUTPUT @ @ PARSE
2 PARSE @ @ PART
2 PARSE @ REPRESENT @
2 PARSE @ SENTENCE @
2 PARSE SENTENCE @ @
2 PRINCIPLE @ @ ZOOM
2 REGARD @ @ WAY
2 REPRESENT @ CONTENT @
2 SEMANTIC @ @ TEXT
2 SENTENCE @ @ CENTRAL
2 SIMPLE PROGRAM @ @
2 SYSTEM @ GENERATE @
2 TEXT @ @ WORD
2 TEXT @ AUTOMATIC MEAN
2 TEXT STRUCTURE @ @
2 WAY @ CONTENT @
2 WORD @ REPRESENT @
2 WORD LIST @ @
2 WORD MEAN @ @
2 WORD SEMANTIC @ @
2 ZOOM CLOSE LOOK @

BILAGA 6 - SAMMANFATTNING 1 (3 059 ORD)

Automatic Abstracting of Contents in Text

Automatic Abstracting of Contents in Text.

It is suggested that CONSTRUCTING A SYNOPSIS OF A GIVEN TEXT BY AUTOMATIC MEANS FROM THE TEXT ITSELF might be done on the basis of a statistical analysis of the vocabulary and collocations in the vocabulary, followed by simple parsing of selected parts of the text, and simple textual analysis, everything based on the results of the vocabulary analysis.

1. A Model for Computerized Generation of Abstracts

2. Vocabulary Analysis

5. Generating Natural Language Texts

1. A Model for Computerized Generation of Abstracts

In part II of this article, the application to a sample text of a model for automatic generation of abstracts from natural language texts will be shown.

To overcome THE FACTORS PREVENTING THIS PICTURE FROM COMING TRUE, we would have to have ways of abstracting the content of texts more or less regardless of the way THE CONTENT is expressed and ABSTRACTING THE CONTENT OF TEXTS MORE OR LESS REGARDLESS OF THE WAY THE CONTENT IS EXPRESSED demands large information systems with immense amounts of "knowledge of the world" in THE INFORMATION SYSTEMS.

One of THE OTHER WAYS would be the abstraction of the content in a text as THE CONTENT is expressed in that text, i.e. giving an abstract of a text, not by means of a general knowledge model, but based on a more surface oriented analysis of the given text.

What I would like to suggest is a research effort aimed at constructing an automatic system for the generation of abstracts of longer, computer readable texts.

To achieve this goal would mean a gathering of knowledge from different fields, such as computational linguistics, text linguistics, and artificial intelligence, as well as a good deal of experimental work on natural language texts.

Still, I would claim that TO ACHIEVE THIS GOAL is a possible project, a project that besides THE PROJECT'S final result would have several highly useful spin-off products, such as parsing mechanisms, thesauri, tools for text structuring, etc.

What would a system generating such surface-based abstracts of texts do, then, and what prerequisites are necessary to get a desired output from THE SYSTEM?

From this structure, THE SYSTEM would generate a readily interpretable representation of the basic content of the original text, preferably in the form of natural language sentences.

To do all THAT IS NECESSARY TO GET A DESIRED OUTPUT FROM THE SYSTEM, it will be necessary to have methods for vocabulary analysis and assessing of relative importance of different words in a text, for parsing sentences to catch relations between words and expressions, for finding overall textual structures, and for generating a natural language output representing the results.

The source text is analysed at three levels: vocabulary, syntax and text.

The processes are highly interdependent; especially the vocabulary analysis will govern the direction and depth of the other two analysing components.

Only sentences containing central concepts will get fully parsed, and textual analysis will concentrate on the interrelations of those concepts and the text structure as reflected by THOSE CONCEPTS.

If the material at hand is found to be insufficient in some respect for generating a coherent abstract of the desired length and informational density, it should be possible to return to the source text with specified demands for more facts.

2. Vocabulary Analysis

In connection with the generation of abstracts, a possible solution to the problem of word meaning is simply to let the words mean "themselves".

Depending on what the analysis is to be used for, TO ADD SOME EXTRA PREREQUISITES TO COVER THE MORE METAPHORICAL USE IN (1B) may

sometimes a necessary step, but for generating an abstract to be read and interpreted by someone who would understand the original text without difficulties, there is not much point in dissecting single words.

Grammatical form words can often be left out except when THE GRAMMATICAL FORM WORDS modify some central concept in an important way

Lists of stop words, i.e. words that can be disregarded at this stage analysis, can easily be constructed by listing words from so-called closed categories, which are mostly form words.

The words that are left after such a reduction are (supposed to be) the content words.

Among the content words, nouns are regarded as carrying most semantic load.

Some interesting experiments have been carried out using computer generated lists of nouns for automatic indexing of texts (Källgren 1984) and people show surprising accuracy in reconstructing the content of a text when given THE TEXT'S nouns in running order.

Adjectives and adverbials can sometimes add or modify content in considerable ways (e.g. negation), which strongly motivates THE ADJECTIVES' AND ADVERBIALS' inclusion among the content words.

First there is the frequency of words within a text, where a more frequent word should give more of the content of the text than A WORD that is less frequent.

That VERY FREQUENT WORDS OFTEN CARRY LESS INFORMATION is TRUE for form words and the neutral verbs mentioned above is easy to see, but there also words that are so frequent in THE WORDS' specific text type as to become more or less devoid of content.

In complement to the general stop list mentioned above, which are common to all texts, one should also consider constructing genre-specific stop lists with content words that are characteristic of a certain text type but so common in such texts as to be almost non-informative (except as a signal of genre).

As a complement to genre-specific lists of stop words, there may however also be certain words that are so important in THE WORDS' specific area that a single mentioning of CERTAIN WORDS deserves notice.

CERTAIN WORDS can then be put on lists of key words and be included in the core vocabulary of a text regardless of how often and where CERTAIN WORDS occur in THE TEXT.

How long is the source text, and how long/short should the desired abstract be?

For short source texts, matters of word frequency within the texts will hardly come into question.

In a large-scale study, Phillips (1985) has shown that collocations of content words which are significantly more frequent than expected can give a fairly good idea of the content of longer scientific texts.

By 'content word', Phillips means about the same as I did above, with important addition that PHILLIPS lemmatizes across category boundaries bringing together words like 'differ', 'different' and 'difference' to one lemma.

In part II of this article, I have made a small and intuitively based collocational analysis by hand to show that this method is something to look deeper into, and that the internal structure of the vocabulary not only the vocabulary itself, contains much useful information about the content of a text.

In order to represent content relationships in text, the latter kind are quite as important to represent.

None of the existing types of parsers seems to be powerful enough for analysis of natural texts about non-trivial worlds in the way that will be necessary in connection with generation of abstracts, but there is still much to be learned from THE EXISTING TYPES OF PARSERS.

The vocabulary analysis sketched above will give hints about which sentences are central and need to be carefully parsed.

Besides THAT THE VOCABULARY ANALYSIS SKETCHED ABOVE WILL GIVE HINTS ABOUT WHICH SENTENCES ARE CENTRAL AND NEED TO BE CAREFULLY PARSED, it is also desirable to have ways of doing partial

parses of sentences, in order trace the relationships of the central terms in the rest of the text.

Text linguistics is a widely varied discipline that seemed very promising in the early seventies.

Text linguistic studies are mostly either purely descriptive, in which case THE TEXT LINGUISTIC STUDIES can give interesting pictures of single texts but can rarely be generalized, or THE TEXT LINGUISTIC STUDIES are highly theoretical at a level where the elaborated theoretical models presented have little or no relevance for analysis of natural texts.

The main reason for this unsatisfactory state seems to be the fact that empirically based text linguistic research is extremely toilsome and unrewarding, with all the intuitively interesting features of text constantly escaping the tools of the analyser.

Form class assignment can be disregarded, as in the vocabulary analysis described above.

The important thing is to maintain the connection with the vocabulary analysis.

When establishing coherence chains in a given text, it is not a question of finding everything that could possibly be related to each word in isolation, it is a question of finding the words of that text that are related to each other in the sense that THE WORDS have in the text.

In longer texts the partitioning in chapters and paragraphs can give useful information, and content words appearing in headlines mostly add more to the textual content than otherwise.

The application of all these different kinds of analysis will leave us with a skeleton of the source text, where some clauses stand out as important, while other parts are blurred down and only noticed as far some central concept appear there.

5. Generating Natural Language Texts

Perhaps the easiest and most concise way of representing the core of the content of a text is to give a list of key words.

In PHILLIPS' study of several longer texts, Phillips (1985) contents himself with presenting structured lists of collocations to give a rough picture of the content of PHILLIPS' texts.

Generation of coherent text from some kind of non-textual database is done in many expert systems and other kinds of question-answering systems.

Information can be lacking because THE INFORMATION was not there in the source text, or because the analysis did not find THE INFORMATION, or underestimated THE INFORMATION'S relevance, or because THE INFORMATION was stored in an erroneous way, or was obscurely expressed in the output.

Lists of key words can be slightly misleading, especially if some important word has been overlooked, but sentences generated from word lists can be totally and utterly wrong and make false statements about the content of a text.

Within predecided limits, which THE LENGTH OF AN ABSTRACT is not allowed to go beyond, the abstract should be free to vary in length according the amount of information in the source text.

Two source texts of the same length can have optimal abstracts of quite different lengths, depending on the density of information in the source texts.

Hopefully, this short overview has given a picture of what there is to be found in different fields of linguistics as concerns the automatic abstraction of content in text.

To give an idea of the possible output from a system constructed along the lines sketched in part I, I have hand-processed a randomly chosen scientific text according to my hypothetical model in the way a computer would do THE PROCESSING, had THE COMPUTER access to methods of the above-mentioned kinds.

This source text consists of 246 words.

(1) Source text

With the aid of a relatively simple program a computer can be converted into a kind of microscope for viewing the boundary of the Mandelbrot set.

In principle one can zoom in for a closer look at any part of the set any magnification.

Approaching the Mandelbrot set, one finds that each wart is a tiny figure shaped much like the parent set.

Zooming in for a close look at one of the tiny figures, however, opens up an entirely different pattern: a riot of organic-looking tendrils and curlicues sweeps out in whorls and rows.

In the center of this center, so to speak, is a four-way bridge with four more curlicues, and in the center of these curlicues another version of the Mandelbrot set is found.

The magnified version is not quite the same Mandelbrot set.

As the zoom continues, such objects seem to reappear, but a closer look always turns up differences.

The first step of the abstracting process will be a vocabulary analysis.

THE VOCABULARY ANALYSIS' aim is to single out the content words of the text.

Form words, which can be easily specified by lists, are taken away, except pronouns which are replaced by THE PRONOUNS' corresponding noun phrase written in capitals.

(My conception of what is to be counted as form words is explained in 2.1 above.)

TAKING AWAY FORM WORDS AND REPLACING PRONOUNS will give Reduced Version 1 with 128 words, a reduction by almost one half of the original size.

(2) Reduced Version 1

Form words and phrases and empty verbs are taken away, pronouns replaced by full forms in capitals

aid relatively simple program computer converted microscope viewing boundary Mandelbrot set.

principle zoom closer look part set magnification.

Approaching Mandelbrot set, finds wart tiny figure shaped much like parent set.

Zooming close look tiny figures, opens entirely different pattern: riot organic-looking tendrils curlicues sweeps whorls rows.

center center, four-way bridge more curlicues, center curlicues version Mandelbrot set found.

magnified version not quite Mandelbrot set.

zoom continues, objects seem reappear, closer look always turns differences.

In the next step, the words of this text are brought together in lemmas with the same root, regardless of THE WORDS' grammatical class.

(3) Content words, lemmatized and alphabetically sorted

In Phillips' model, the collocates of a content word are all other content words appearing within a span from four words of running text before an occurrence of the word till four words after THE WORD.

The span is thus established in the text, (1) above, counting also the words that were removed in the first step, although the only words qualify as collocates are content words from the list (3).

For a collocation of two or more words to be regarded as substantially contributing to the content of the text THE COLLOCATION has to have a certain frequency in the text.

Mandelbrot set find

close look zoom

tiny figure wart

Thus, the word 'Mandelbrot' only occurs immediately in front of 'set', forming a semantic unity, while the word 'set' can also appear alone or in connection with other content words.

'Mandelbrot set' and 'find' seem to have no natural connection, while 'close look' and 'zoom' are felt as highly interrelated.

Now we can construct Reduced Version 2, which is simply built up by the sentences of the source text that contain members from two or more of collocations listed above.

What we would like to do here would be to parse the concerned sentences carefully, represent the relations between the concepts that have been found to be

of interest, i.e. the significantly collocating words, and then generate a new text from this basic content.

Reduced Version 2 has the form words from the source text restored and contains 107 words, about 40% of the source text.

(6a) Reduced Version 2

In principle one can zoom in for a closer look at any part of the MANDELBROT SET at any magnification.

Approaching the Mandelbrot set, one finds that each wart is a tiny figure shaped much like the parent set.

Zooming in for a closer look at one of the tiny figures, however, opens up an entirely different pattern.

In the center of these curlicues another version of the Mandelbrot set is found.

The magnified version is not quite the same Mandelbrot set.

Partly THE FACT THAT ABSTRACT (6A) IS A BIT TOO LONG IN RELATION TO THE SOURCE is due to the shortness of the source text, a short text cannot be abbreviated as much as a longer text.

Treating the text the way a careful human editor would do gives Reduced Version 3, which at best may be possible to reach also by some developed form of computerized Writing Assistant (Richardson 1985).

The number of words in THE TEXT is 86.

(6b) Reduced Version 3

One can zoom in for a closer look at any part of the MANDELBROT SET at any magnification.

Each wart is a tiny figure shaped much like the parent set.

Zooming in for a close look at one of the tiny figures opens up a different pattern.

In the center of these curlicues another version of the Mandelbrot set is found.

The magnified version is not the same Mandelbrot set.

Still THE VERSION is not good enough for a final result, but THE VERSION is good enough as a basis for more refined processing making use of semantic relations, careful parsing, and text structural information along the lines mentioned above and in part I: Discussion of Relevant Research.

What is important to notice is that all information necessary for deriving a real good synopsis of the source text (1) is in fact present in the short versions in (6), so a more detailed and elaborate analysis can be restricted to an abridged version.

THAT A MORE DETAILED AND ELABORATE ANALYSIS CAN BE RESTRICTED TO AN ABRIDGED VERSION will make the processing quicker and easier with less risk of getting lost in all the material of the source text, while the source is still there as a miniature database to look for facts in if need be.

Analysis of the abridged version in a way that is similar to "interpreting" and "understanding" THE ABRIDGED VERSION and then restructuring THE ABRIDGED VERSION by means of a natural language generator would at last give a fully acceptable output, perhaps something like (7).

(7) Any part of the Mandelbrot set can be magnified.

The magnified version will be a figure much like the first Mandelbrot set but not quite the same.

There exists knowledge about these matters in many fields, and if that knowledge is gathered in an effort at automatic abstraction of content, new knowledge will quickly be derived.

Hendrix, Gary G. et al. 1978. Developing a Natural Language Interface to Complex Data. ACM Transactions on Database Systems 8 (3).

King, Margaret 1983. Parsing Natural Language. London: Academic Press.

Lehnert, Wendy G. & Ringle, Martin H. (eds.) 1982. Strategies for Natural Language Processing. Hillsdale, N.J.: Lawrence Erlbaum Ass.

In Walker, Donald E. et al. (eds.): Natural Language in Information Science. Stockholm: Skriptor.

Phillips, Martin 1985. Aspects of Text Structure. Amsterdam: Elsevier.

Richardson, Stephen 1985. CRITIQUE: Enhanced Text Critiquing Using a Natural Language Parser.

Proceedings from ELS Conference on Natural Language Applications, September 10-12, 1985. IBM Denmark.

Sparck Jones, Karen & Wilks, Yorick (eds.) 1983. Automatic Natural Language Parsing. Chichester: Horwood.

Winograd, Terry 1972. Understanding Natural Language. New York: Academic Press.

BILAGA 7 - SAMMANFATTNING 2 (2 136 ORD)

It is suggested that CONSTRUCTING A SYNOPSIS OF A GIVEN TEXT BY AUTOMATIC MEANS FROM THE TEXT ITSELF might be done on the basis of a statistical analysis of the vocabulary and collocations in the vocabulary, followed by simple parsing of selected parts of the text, and simple textual analysis, everything based on the results of the vocabulary analysis.

2. Vocabulary Analysis

5. Generating Natural Language Texts

In part II of this article, the application to a sample text of a model for automatic generation of abstracts from natural language texts will be shown.

One of THE OTHER WAYS would be the abstraction of the content in a text as THE CONTENT is expressed in that text, i.e. giving an abstract of a text, not by means of a general knowledge model, but based on a more surface oriented analysis of the given text.

To achieve this goal would mean a gathering of knowledge from different fields, such as computational linguistics, text linguistics, and artificial intelligence, as well as a good deal of experimental work on natural language texts.

From this structure, THE SYSTEM would generate a readily interpretable representation of the basic content of the original text, preferably in the form of natural language sentences.

To do all THAT IS NECESSARY TO GET A DESIRED OUTPUT FROM THE SYSTEM, it will be necessary to have methods for vocabulary analysis and assessing of relative importance of different words in a text, for parsing sentences to catch relations between words and expressions, for finding overall textual structures, and for generating a natural language output representing the results.

The source text is analysed at three levels: vocabulary, syntax and text.

The processes are highly interdependent; especially the vocabulary analysis will govern the direction and depth of the other two analysing components.

If the material at hand is found to be insufficient in some respect for generating a coherent abstract of the desired length and informational density, it should be possible to return to the source text with specified demands for more facts.

2. Vocabulary Analysis

The words that are left after such a reduction are (supposed to be) the content words.

Among the content words, nouns are regarded as carrying most semantic load.

Some interesting experiments have been carried out using computer generated lists of nouns for automatic indexing of texts (KÑllgren 1984) and people show surprising accuracy in reconstructing the content of a text when given THE TEXT'S nouns in running order.

Adjectives and adverbials can sometimes add or modify content in considerable ways (e.g. negation), which strongly motivates THE ADJECTIVES' AND ADVERBIALS' inclusion among the content words.

First there is the frequency of words within a text, where a more frequent word should give more of the content of the text than A WORD that is less frequent.

In complement to the general stop list mentioned above, which are common to all texts, one should also consider constructing genre-specific stop lists with content words that are characteristic of a certain text type but so common in such texts as to be almost non-informative (except as a signal of genre).

How long is the source text, and how long/short should the desired abstract be?

For short source texts, matters of word frequency within the texts will hardly come into question.

In a large-scale study, Phillips (1985) has shown that collocations of content words which are significantly more frequent than expected can give a fairly good idea of the content of longer scientific texts.

By 'content word', Phillips means about the same as I did above, with important addition that PHILLIPS lemmatizes across category boundaries bringing together words like 'differ', 'different' and 'difference' to one lemma.

In part II of this article, I have made a small and intuitively based collocational analysis by hand to show that this method is something to look deeper into, and that the internal structure of the vocabulary not only the vocabulary itself, contains much useful information about the content of a text.

In order to represent content relationships in text, the latter kind are quite as important to represent.

The vocabulary analysis sketched above will give hints about which sentences are central and need to be carefully parsed.

Besides THAT THE VOCABULARY ANALYSIS SKETCHED ABOVE WILL GIVE HINTS ABOUT WHICH SENTENCES ARE CENTRAL AND NEED TO BE CAREFULLY PARSED, it is also desirable to have ways of doing partial parses of sentences, in order trace the relationships of the central terms in the rest of the text.

Form class assignment can be disregarded, as in the vocabulary analysis described above.

The important thing is to maintain the connection with the vocabulary analysis.

When establishing coherence chains in a given text, it is not a question of finding everything that could possibly be related to each word in isolation, it is a question of finding the words of that text that are related to each other in the sense that THE WORDS have in the text.

In longer texts the partitioning in chapters and paragraphs can give useful information, and content words appearing in headlines mostly add more to the textual content than otherwise.

The application of all these different kinds of analysis will leave us with a skeleton of the source text, where some clauses stand out as important, while other parts are blurred down and only noticed as far some central concept appear there.

5. Generating Natural Language Texts

Perhaps the easiest and most concise way of representing the core of the content of a text is to give a list of key words.

In PHILLIPS' study of several longer texts, Phillips (1985) contents himself with presenting structured lists of collocations to give a rough picture of the content of PHILLIPS' texts.

Generation of coherent text from some kind of non-textual database is done in many expert systems and other kinds of question-answering systems.

Information can be lacking because THE INFORMATION was not there in the source text, or because the analysis did not find THE INFORMATION, or underestimated THE INFORMATION'S relevance, or because THE INFORMATION was stored in an erroneous way, or was obscurely expressed in the output.

Lists of key words can be slightly misleading, especially if some important word has been overlooked, but sentences generated from word lists can be totally and utterly wrong and make false statements about the content of a text.

Within predecided limits, which THE LENGTH OF AN ABSTRACT is not allowed to go beyond, the abstract should be free to vary in length according the amount of information in the source text.

Two source texts of the same length can have optimal abstracts of quite different lengths, depending on the density of information in the source texts.

This source text consists of 246 words.

(1) Source text

With the aid of a relatively simple program a computer can be converted into a kind of microscope for viewing the boundary of the Mandelbrot set.

In principle one can zoom in for a closer look at any part of the set any magnification.

Approaching the Mandelbrot set, one finds that each wart is a tiny figure shaped much like the parent set.

Zooming in for a close look at one of the tiny figures, however, opens up an entirely different pattern: a riot of organic-looking tendrils and curlicues sweeps out in whorls and rows.

In the center of this center, so to speak, is a four-way bridge with four more curlicues, and in the center of these curlicues another version of the Mandelbrot set is found.

The magnified version is not quite the same Mandelbrot set.

As the zoom continues, such objects seem to reappear, but a closer look always turns up differences.

The first step of the abstracting process will be a vocabulary analysis.

THE VOCABULARY ANALYSIS' aim is to single out the content words of the text.

aid relatively simple program computer converted microscope viewing boundary Mandelbrot set.

principle zoom closer look part set magnification.

Approaching Mandelbrot set, finds wart tiny figure shaped much like parent set.

Zooming close look tiny figures, opens entirely different pattern: riot organic-looking tendrils curlicues sweeps whorls rows.

center center, four-way bridge more curlicues, center curlicues version Mandelbrot set found.

magnified version not quite Mandelbrot set.

zoom continues, objects seem reappear, closer look always turns differences.

In the next step, the words of this text are brought together in lemmas with the same root, regardless of THE WORDS' grammatical class.

(3) Content words, lemmatized and alphabetically sorted

In Phillips' model, the collocates of a content word are all other content words appearing within a span from four words of running text before an occurrence of the word till four words after THE WORD.

The span is thus established in the text, (1) above, counting also the words that were removed in the first step, although the only words qualify as collocates are content words from the list (3).

For a collocation of two or more words to be regarded as substantially contributing to the content of the text THE COLLOCATION has to have a certain frequency in the text.

Mandelbrot set find

close look zoom

Thus, the word 'Mandelbrot' only occurs immediately in front of 'set', forming a semantic unity, while the word 'set' can also appear alone or in connection with other content words.

'Mandelbrot set' and 'find' seem to have no natural connection, while 'close look' and 'zoom' are felt as highly interrelated.

Now we can construct Reduced Version 2, which is simply built up by the sentences of the source text that contain members from two or more of collocations listed above.

What we would like to do here would be to parse the concerned sentences carefully, represent the relations between the concepts that have been found to be of interest, i.e. the significantly collocating words, and then generate a new text from this basic content.

Reduced Version 2 has the form words from the source text restored and contains 107 words, about 40% of the source text.

In principle one can zoom in for a closer look at any part of the MANDELBROT SET at any magnification.

Approaching the Mandelbrot set, one finds that each wart is a tiny figure shaped much like the parent set.

Zooming in for a closer look at one of the tiny figures, however, opens up an entirely different pattern.

In the center of these curlicues another version of the Mandelbrot set is found.

The magnified version is not quite the same Mandelbrot set.

Partly THE FACT THAT ABSTRACT (6A) IS A BIT TOO LONG IN RELATION TO THE SOURCE is due to the shortness of the source text, a short text cannot be abbreviated as much as a longer text.

The number of words in THE TEXT is 86.

One can zoom in for a closer look at any part of the MANDELBROT SET at any magnification.

Zooming in for a close look at one of the tiny figures opens up a different pattern.

In the center of these curlicues another version of the Mandelbrot set is found.

The magnified version is not the same Mandelbrot set.

What is important to notice is that all information necessary for deriving a real good synopsis of the source text (1) is in fact present in the short versions in (6), so a more detailed and elaborate analysis can be restricted to an abridged version.

THAT A MORE DETAILED AND ELABORATE ANALYSIS CAN BE RESTRICTED TO AN ABRIDGED VERSION will make the processing quicker and easier with less risk of getting lost in all the material of the source text, while the source is still there as a miniature database to look for facts in if need be.

Analysis of the abridged version in a way that is similar to "interpreting" and "understanding" THE ABRIDGED VERSION and then restructuring THE ABRIDGED VERSION by means of a natural language generator would at last give a fully acceptable output, perhaps something like (7).

(7) Any part of the Mandelbrot set can be magnified.

The magnified version will be a figure much like the first Mandelbrot set but not quite the same.

Hendrix, Gary G. et al. 1978. Developing a Natural Language Interface to Complex Data. ACM Transactions on Database Systems 8 (3).

King, Margaret 1983. Parsing Natural Language. London: Academic Press.

Lehnert, Wendy G. & Ringle, Martin H. (eds.) 1982. Strategies for Natural Language Processing. Hillsdale, N.J.: Lawrence Erlbaum Ass.

In Walker, Donald E. et al. (eds.): Natural Language in Information Science. Stockholm: Skriptor.

Richardson, Stephen 1985. CRITIQUE: Enhanced Text Critiquing Using a Natural Language Parser.

Proceedings from ELS Conference on Natural Language Applications, September 10-12, 1985. IBM Denmark.

Sparck Jones, Karen & Wilks, Yorick (eds.) 1983. Automatic Natural Language Parsing. Chichester: Horwood.

Winograd, Terry 1972. Understanding Natural Language. New York: Academic Press.

BILAGA 8 - SAMMANFATTNING 3 (991 ORD)

Automatic Abstracting of Contents in Text

Automatic Abstracting of Contents in Text.

To overcome THE FACTORS PREVENTING THIS PICTURE FROM COMING TRUE, we would have to have ways of abstracting the content of texts more or less regardless of the way THE CONTENT is expressed and ABSTRACTING THE CONTENT OF TEXTS MORE OR LESS REGARDLESS OF THE WAY THE CONTENT IS EXPRESSED demands large information systems with immense amounts of "knowledge of the world" in THE INFORMATION SYSTEMS.

One of THE OTHER WAYS would be the abstraction of the content in a text as THE CONTENT is expressed in that text, i.e. giving an abstract of a text, not by means of a general knowledge model, but based on a more surface oriented analysis of the given text.

To do all THAT IS NECESSARY TO GET A DESIRED OUTPUT FROM THE SYSTEM, it will be necessary to have methods for vocabulary analysis and assessing of relative importance of different words in a text, for parsing sentences to catch relations between words and expressions, for finding overall textual structures, and for generating a natural language output representing the results.

Word semantics, working with the meanings of words and THE WORDS' interrelations, is an old and well-established branch of linguistics that every now and then receives an increased interest from more logico-semantically oriented movements.

Texts get THE TEXTS' meaning from the words occurring in THE TEXTS, and words get THE WORDS' precise meaning from the (con-)text where THE WORDS occur.

The words that are left after such a reduction are (supposed to be) the content words.

Among the content words, nouns are regarded as carrying most semantic load.

Some interesting experiments have been carried out using computer generated lists of nouns for automatic indexing of texts (KÑllgren 1984) and people show surprising accuracy in reconstructing the content of a text when given THE TEXT'S nouns in running order.

Adjectives and adverbials can sometimes add or modify content in considerable ways (e.g. negation), which strongly motivates THE ADJECTIVES' AND ADVERBIALS' inclusion among the content words.

First there is the frequency of words within a text, where a more frequent word should give more of the content of the text than A WORD that is less frequent.

In complement to the general stop list mentioned above, which are common to all texts, one should also consider constructing genre-specific stop lists with content words that are characteristic of a certain text type but so common in such texts as to be almost non-informative (except as a signal of genre).

In a large-scale study, Phillips (1985) has shown that collocations of content words which are significantly more frequent than expected can give a fairly good idea of the content of longer scientific texts.

By 'content word', Phillips means about the same as I did above, with important addition that PHILLIPS lemmatizes across category boundaries bringing together words like 'differ', 'different' and 'difference' to one lemma.

In part II of this article, I have made a small and intuitively based collocational analysis by hand to show that this method is something to look deeper into, and that the internal structure of the vocabulary not only the vocabulary itself, contains much useful information about the content of a text.

In order to represent content relationships in text, the latter kind are quite as important to represent.

When establishing coherence chains in a given text, it is not a question of finding everything that could possibly be related to each word in isolation, it is a question of finding the words of that text that are related to each other in the sense that THE WORDS have in the text.

In longer texts the partitioning in chapters and paragraphs can give useful information, and content words appearing in headlines mostly add more to the textual content than otherwise.

Perhaps the easiest and most concise way of representing the core of the content of a text is to give a list of key words.

In PHILLIPS' study of several longer texts, Phillips (1985) contents himself with presenting structured lists of collocations to give a rough picture of the content of PHILLIPS' texts.

TO PRESENT STRUCTURED LISTS OF COLLOCATIONS TO GIVE A ROUGH PICTURE OF THE CONTENT OF TEXTS may be useful for e.g. automatic indexing, where information retrieval system is to search among the key concepts and to present the full text in the case of a hit.

Lists of key words can be slightly misleading, especially if some important word has been overlooked, but sentences generated from word lists can be totally and utterly wrong and make false statements about the content of a text.

Hopefully, this short overview has given a picture of what there is to be found in different fields of linguistics as concerns the automatic abstraction of content in text.

THE VOCABULARY ANALYSIS' aim is to single out the content words of the text.

In the next step, the words of this text are brought together in lemmas with the same root, regardless of THE WORDS' grammatical class.

(3) Content words, lemmatized and alphabetically sorted

In Phillips' model, the collocates of a content word are all other content words appearing within a span from four words of running text before an occurrence of the word till four words after THE WORD.

The span is thus established in the text, (1) above, counting also the words that were removed in the first step, although the only words qualify as collocates are content words from the list (3).

For a collocation of two or more words to be regarded as substantially contributing to the content of the text THE COLLOCATION has to have a certain frequency in the text.

Thus, the word 'Mandelbrot' only occurs immediately in front of 'set', forming a semantic unity, while the word 'set' can also appear alone or in connection with other content words.

The number of words in THE TEXT is 86.