LINKÖPINS UNIVERSITET

SimSum

En studie om automatisk sammanfattning och omskrivning av texter

Sammanfattning

Anton Jeppsson
Samuel Johnson
Erik Karlsson
Christofer Malmberg
Victor Sjölin
Åsa Svensson

2012-05-31

Sammanfattning

Arbetet som presenteras i den här rapporten är resultatet av ett antal experiment där programmet SimSum testas. Rapportens fokus ligger på att utvärdera två möjliga sätt att sammanfatta texter automatiskt och undersöka om det ena sättet ger bättre resultat än det andra. Rapporten och undersökningen är en del i ett större projekt som kallas Easy Reader, vars syfte är att ta fram ett program som kan underlätta läsandet av texter för människor med lässvårigheter.

Undersökningen har utförts genom att försöksdeltagare har fått läsa två olika texter, förenklade med varsin metod, och sedan fått svara på frågor som testat deras förståelse av textens innehåll. Efter detta har de sedan fått betygsätta texten genom ett antal frågor rörande textens läsbarhet och så har även läsbarhetsmått utvärderats.

Resultatet av undersökningen visar att det inte finns någon signifikant skillnad i försöksdeltagarnas förmåga att svara på frågorna, oavsett förenklingsmetod. Vi finner dock en signifikant skillnad i antal ord mellan texterna och drar därför slutsatsen att det är bättre att först skriva om texten och sedan sammanfatta den. Eftersom antal rätta svar inte skiljer sig tolkar vi det som att omskrivning först ger en effektivare förenkling av en text.

Inledning

En stor del av den information som vi är i behov av i vårt dagliga liv finns bara tillgänglig som skriven text och för de personer som lider av någon form av lässvårighet kan detta innebära stora problem. Programmet som utvärderas i den här rapporten är tänkt att underlätta läsandet av olika sorters texter. Programmet är utformat för att göra texter mer lätttillgängliga för människor med lässvårigheter genom att förkorta och skriva om texterna, utan att ta bort viktig information. Målet med den här rapporten är att avgöra om det finns en skillnad i läsbarheten hos de texter som programmet producerar beroende på i vilken ordning de två olika modulerna som behandlar texten körs.

Syfte

En text kan till exempel förenklas genom sammanfattning och omskrivning. Syftet med denna rapport var att ta reda på i vilken ordning detta bör ske. Hypotesen är att sammanfattning bör användas före omskrivning på grund av att omskrivningsregler kan påverka sammanfattningsprocessen negativt.

Bakgrund

Programmet vi använder oss av kallas SimSum och består av två moduler; en simplifierare (CogFLUX) som skriver om en text till enkel svenska, benäms även som O i denna rapport, och en summerare (CogSUM) som sammanfattar en text genom att välja ut de mest relevanta meningarna, benämns i denna rapport även som S. I de fall vi nämner SO innebär det att texten först är sammanfattad och sedan omskriven. Det omvända gäller vid OS.

SimSum är specifikt utvecklat för denna rapport av Robin Keskisärkkä. Texterna som skapas har en förutbestämd sammanfattningsgrad och använder ett bestämt antal omskrivningsregler. Skillnaden i de producerade texterna är således enbart i vilken ordning som modulerna körs. Programmet räknar även automatiskt ut nio olika former av läsbarhetsmått. För en exempelkörning se bild 1.

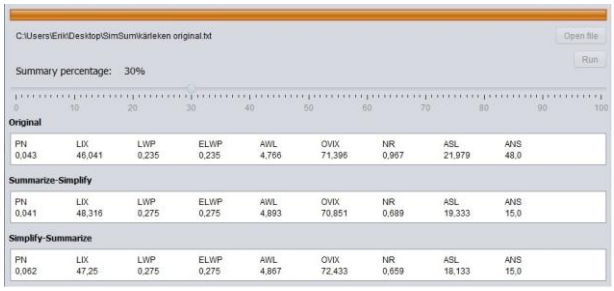

Bild 1. SimSum

CogSUM

CogSUM är namnet på ett program som skapades av Axelsson et al. (2008). CogSUM är en extraktionsbaserad sammanfattare som använder sig av RandomIndex samt en svagt modifierad variant av PageRank. Att sammanfattningen är extraktionsbaserad innebär att man aldrig skriver om några meningar utan enbart väljer ut, det vill säga extraherar, de viktigaste meningarna i dokumentet för att sammanfatta det.

CogSUM börjar med att läsa in hela texten och skapa en kontextvektor för varje ord i sin lemmatiserade form. Kontextvektorerna bygger på RandomIndex vilket innebär att man först slumpar ut en indexvektor till varje ord. En indexvektor kan ses som en unik kod som man ger varje ord och som har ett begränsat antal dimensioner (eller element). Med hjälp av indexvektorerna kan man sedan skapa kontextvektorer genom att lägga ihop två eller flera indexvektorer.

Kontextvektorn viktas sedan med hjälp av ett viktningsschema. Detta schema kan liknas vid en av ovanstående vektorer och används för att kunna påverka så att inte alla ord i kontextvektorn får lika hög grad av betydelse. Desto längre ifrån fokusordet står ett annat ord desto mindre viktning får det. Sedan läggs alla dessa kontextvektorer ihop och delas med det totala antalet vektorer för att få en dokumentvektor. Den dokumentvektorn används som en mall för innehållet i texten och för varje mening räknas sedan en meningsvektor ut på samma sätt som för dokumentvektorn. Genom att sedan göra en cosinus beräkning får man fram skillnaden mellan vektorerna.

För att sedan få fram de viktigaste meningarna används PageRank. Det har fått sitt namn efter uppfinnaren Lawrence Page (Page, Brin, & Motwani, 1998) och användes ursprungligen av Google för att förutspå hur en användare navigerar webben. Page förutsåg att en användare startar vid en slumpmässig nod (webbsida) och navigerar sedan från den via länkar tills användaren tröttnar och börjar vid en ny, slumpmässig nod. Sannolikheten för att en användare besöker en sida X är dess PageRank.

I det här fallet är noderna meningarna från texten som ska sammanfattas, istället för webbsidorna som i PageRank ekvationen. Dessa noder pekar sedan på andra noder och detta utnyttjar man för att få fram de viktigaste meningarna. Det blir helt enkelt så att varje mening "röstar" på alla andra meningar den hör ihop med och de meningar som får flest röster är de meningarna som CogSUM sedan väljer ut. I CogSUM används de första cosinusberäkningarna som startvärden i PageRank-uträkningen för att så snabbt som möjligt finna det optimala värdet.

CogFLUX

CogFLUX är en omskrivningsmodul gjord av Rybing, Smith, & Silvervarg (2010). Syftet med CogFLUX är att hjälpa den stora del av den vuxna befolkningen som har problem med att läsa på grundskolenivå genom att automatiskt skriva om texter till lättläst svenska. CogFLUX skriver om meningarna utifrån Deckers (2003) omskrivningsregler och använder hennes syntax och förslag till omskrivningsregler.
CogFLUX är uppbyggt kring tre huvudprocesser som i sin tur är uppbyggda av två eller tre mindre processer: PreProcessor, TransformationsProcessor och PostProcessor.

I preprocessorn återfinns en rad olika moduler. Alla dessa moduler har till uppgift att bearbeta texten så att den blir så enkel som möjligt att utföra omskrivningen på. Detta innebär att ordklasstagga alla ord samt att dela upp meningarna i frasstrukturträd eftersom Decker (2003) regler är på det formatet. Nästa stora steg för meningen är att gå över till transformationsprocessorn. Här återfinns alla moduler som på något sätt har med förenklingen av texten att göra. Sista steget för meningen är att gå igenom postprocessorfasen. Denna fas består av tre moduler som har till uppgift att rensa texten från taggarna och återställa texten i läsvänlig form.

Deckerreglerna

Decker (2003) undersöker manuell omskrivning till lättlästa texter för att se om det finns ett mönster i hur manuell omskrivning av en text görs. Hon undersöker även om det kan användas i ett automatiserat system för textförenkling. Decker tog fram 25 förenklingsregler utifrån de grammatiska skillnaderna som fanns mellan manuellt omskriva texter till lättläst svenska och deras ursprungstexter. Deckers studie riktar sig mer mot lättläst svenska som andra språk och inte åt lättläst svenska för de med svenska som modersmål men som lider av lässvårigheter (dyselexi, ADHD, utvecklingstörningar, etc.) och studien använde sig av Invandrartidningen som plattform, en nyhetstidning riktad mot invandrare.

Metod

För att uppnå vårt syfte har vi använt oss av både kvalitativa och kvantitativa metoder. För att utvärdera de omskrivna och sammanfattade texterna så använde vi oss av försöksdeltagare som fick läsa två texter vardera, svara på frågor på texterna och även på frågor om hur det upplevde texterna. Ett pilottest utfördes för att försäkra att frågorna uppfattades rätt och var till användning för studien. Texterna har även utvärderats utifrån olika läsbarhetsmått så som LIX, OVIX och NR.

Val av deckerregler

CogFLUX skriver om meningarna utifrån en samling deckerregler. Användning av samtliga 13 regler som implementerats i CogFLUX, av Deckers (2003) usprungliga 25, resulterade det i att texten ansågs vara oförståelig. Därför tog vi bort några av reglerna för att öka läsligheten och detta resulterade i att vi använde oss av sex olika omskrivningsregler. Eftersom vi inte fann någon tidigare forskning inom området valde vi regler som gav en blandning av minskad textmassa och behållning av information.

Undersökning med hjälp av försöksdeltagare

Undersökningen var en mixed design, där jämförelsen mellan OS och SO valdes som inomgruppsdesign och jämförelsen med sammanfattningsprocent är en mellangruppsdesign. Undersökningen utfördes genom att varje försöksdeltagare fick läsa två texter. Vilka texter som lästes och i vilken ordning bestämdes i förväg. Eftersom OS och SO jämfördes med en inomgruppsdesign fick varje försöksdeltagare läsa en text av varje sort, det vill säga en OS text och en SO text. Vidare bestämdes graden av sammanfattningsprocent som försöksdeltagaren skulle läsa, antingen 30%, 50% eller 70%. Detta för att kunna undersöka om sammanfattningsprocenten spelade någon roll. Denna uppdelning resulterade i tre olika grupper och inom varje grupp delades försökspersonerna upp så hälften fick läsa OS först och hälften SO först. Inom dessa mindre grupper delades även försöksdeltagarna upp i vilken av texterna de läste först, “Kärlekens Makt och Tårar” eller “Rätt Kost vid Diabetes”. Vilken text som lästes först och vilket

omskrivningssätt denna hade varierades för att undvika eventuella felmarginaler.

Resultat

Alla resultat är beräknade med parade t-test i IBM SPSS Statistics 19. Signifikanta värden (P < 0,05) har markerats med fet stil i tabellerna.

I tabell 1 visas antal rätta svar. Som synes finns inget signifikant resultat (P < 0,05), oavsett procentsats.

Tabell 1. Antal rätt. M står för medelvärde.

	Alla		30%		50%		70%	
	M	P	M	P	M	P	M	P
OS	1.98	0.82	1.77	0.27	2.00	0.08	2.17	1.00
SO	1.94		2.03		1.63		2.17	

Tabell 2 visar medelvärde och sannolikhetsresultat för antal ord beroende på omskrivningsmodul. Notera att texterna som skrivs om med OS är signifikant kortare (t(5)=-2.781 p<0.05).

Tabell 2. Antal ord i texterna

	Alla	
	Medelvärde	P
Antal ord OS	407.6667	**0.039**
Antal ord SO	453.1667	

Slutsats

Med ovanstående resultat anser vi att det alltid är bättre att skriva om en text innan man sammanfattar den, om målet är att få en så kort och innehållsrik text som möjligt. Detta eftersom att omskrivning av meningar leder till att skillnaderna mellan dem minskar. Detta i sin tur underlättar för sammanfattningsmodulen, eller närmare bestämt PageRank, som kommer anse att fler meningar innehåller samma fakta. Eftersom PageRank sedan bara väljer ut en av dessa "sammanklumpade" meningar kommer detta garantera fler unika meningar i sammanfattningen, det vill säga mer unik information. Detta sker dock på bekostnad av flytet i texten.

Om flytet i texten är det viktigaste anser vi att enbart sammanfattningsmodulen bör användas. Detta för att manuellt skrivna texter förhoppningsvis har det optimala flytet samt att i vissa fall, så som skönlitterära texter, ligger poängen i att behålla författarens ursprungsord.

Vi anser också att i vissa fall kan det räcka med att bara använda omskrivningsmodulen (en slutgiltig version med till exempel synonymhantering etc.). Ibland vill man vara säker på att få all information men har till exempel svårt för akademiska texter. För hur man än sammanfattar en text så kommer information alltid att gå förlorad, därför är enbart omskrivning i speciella fall att föredra.

Referenser

Axelsson, M., Bergenholm, E., Carlsson, B., Dahlbom, G., Rybing, J., & Smith, C. (2008). *CogSum - Ett försök att med dagens automatiska informationsextraheringsmetoder och rankningsalgoritmer skapa sammanfattningar i skumläsningssyfte*. Linköpings Universitet.

Decker, A. (2003). *Towards automatic grammatical simplification of Swedish text*. Stockholms Universitet.

Page, L., Brin, S., & Motwani, R. (1998). The PageRank citation ranking: Bringing order to the web. *World Wide Web Internet And Web Information Systems*, *54*(1999-66), 1-17. doi:10.1.1.31.1768

Rybing, J., Smith, C., & Silvervarg, A. (2010). *CogFLUX*. Linköpings universitet.