Automatisk text-sammanfattning

Språkteknologi – mänskliga språk och datorer,
Martin Duneld

Textsammanfattning

- Att extrahera kärnan, essensen, av en text och presentera den i en kortare form med så liten förlust som möjligt när det gäller förmedlad information

Automatisk textsammanfattning

- Automatisk textsammanfattning är en teknik där ett datorprogram sammanfattar en text
- Programmet ges en text och returnerar en **kortare**, förhoppningsvis **icke-redundant** text
- De tidigaste systemen är från 60-talet
- Några kända milstolpar
  - Luhn 1959, Edmunson 1969, Salton 1989

Då och nu

- Då:
  - Processorkraft och digitalt lagringsutrymme var dyrt och litet
  - Lösning: Sammanfatta texter före indexering
- Nu:
  - I och med internets genomslag så har nytt intresse och nya användningsområden uppstått
  - Idag produceras mer text än vad någon kan ta till sig – inom nästan varje tänkbart ämne!
  - Dagens datorer är kraftfulla nog att sammanfatta stora mängder text nästan ögonblickligen

Typer av sammanfattningar

- Indikativa vs informativa
  - Används för snabb kategorisering kontra för att ta till sig innehållet
- Extrakt vs abstrakt
  - Listar fragment av text kontra omfraserar innehållet på ett sammanhängande sätt
- Generiska vs frågefokuserade (”query oriented”)
  - Ger författarens bild kontra speglar användarens intresse
- Bakgrund vs enbart-nyheterna
  - Förutsätter att läsarens förkunskaper är dåliga kontra uppdaterade
- Ett dokument vs flera dokument som källa
  - Bygger på en text kontra smälter samman många texter

Textabstraktion

- Vad vi människor gör
  - Vi läser en text, omtolkar den och skriver om den med våra egna ord
- Med en dator
  - **Semantisk** parsning
  - Översättning till en **formell** representation
  - En uppsättning val gällande vad som skall förmedlas baserat på den formella beskrivningen
  - Textgenerering
    - Nya **syntaktiska** strukturer
    - Nya **lexikala** val

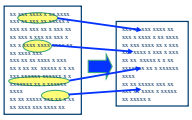
Textextraktion

- Identifiera textens/texternas ämne
- Hitta centrala delar i texten som behandlar ämnet
  - Använd **lingvistiska**, **statistiska** och **heuristiska** metoder för att ranka dessa delar
- Extrahera de högst rankade delarna i texten
  - Till exempel stycken, meningar, satser
- Konkatenera de extraherade delarna i lämplig ordning och presentera denna text

- De allra flesta textsammanfattningssystem bygger på extraktion!

Dator vs människa

- Abstrakt till extrakt (Marcu 98)
  - 10 texter med motsvarande abstrakt skrivna av människor
  - Bad 14 bedömare plocka ut motsvarande stycken ut originaltexterna för att täcka samma innehåll
  - Jämförde abstrakt- och extraktlängd räknat i ord
  - **Extraktlängd = 2.76 * abstraktlängd !!**



Varför då ATS?

- Automatisk textsammanfattning är ännu långt ifrån vad professionella mänskliga sammanfattare presterar, och blir eventuellt aldrig lika bra

- MEN, det är mycket snabbare och billigare!

Metoder för textextraktion I

Rankningskriterier använda i olika extraktionsbaserade sammanfattningssystem:

- **Baseline**: Meningarnas ordning i originaltexten indikerar hur viktiga de är. Första meningen rankar högst, sista lägst.

- **Första meningen**: Första meningen i varje stycke får en högre rank.

Metoder för textextraktion II

- **Genreposition**: Kommer från observationen att vissa genrer placerar viktig information på specifika platser. Nyhetstext placerar den i de inledande meningarna, vetenskaplig text i början och i slutet.

- **Rubriker**: Meningar som innehåller ord som förekommer i rubriker rankas upp.

- **Termfrekvens (tf)**: Innehållsord som är högfrekventa i texten är viktigare än mindre frekventa. Innehållsord tillhör de öppna ordklasserna.

Metoder för textextraktion III

- **Average lexical connectivity**: Antal innehållsord som delas med andra meningar. Hypotesen är att meningar som delar ord med många andra meningar är centrala meningar.

- **Query signature**: Informationsbehovet uttrycks som en sökfråga. Ord i sökfrågan viktas upp i texten.

- **Meningslängd**: Meningslängden antyder vilka meningar som är viktiga.

Metoder för textextraktion IV

- **Triggerord**: Fördefinierad lista på viktiga ord, t.ex. inom en viss domän. Meningar som innehåller dessa ord viktas upp.
- **Numeriska data**: Meningar som innehåller värden, intervall, datum etc. viktas upp.
- **Egennamn**: Meningar som innehåller namn på personer, platser, företag etc. ges högre vikt.

Metoder för textextraktion V

- **Veckodagar och månader**: Meningar som förankrar något i tiden anses centrala.
- **Pronomen och adjektiv**: Adjektiv beskriver, pronomen knyter samman meningar.
- Enkel **kombinationsfunktion**:

$$vikt(S) = \alpha.rubrik(S) + \beta.frekvens(S) + \gamma.triggerord(S) + \delta.position(S)$$

  - Parametrarna optimeras med träningsdata eller experimentellt

Domänord

Med en korpus kan vi utnyttja kunskap om ordens distribution

- ***tf* = termfrekvens**: antal gånger ett ord förekommer i ett dokument
- ***idf* = invers dokumentfrekvens**: antal dokument som ordet förekommer i delat med totala antalet dokument
- ***tf*idf*** indikerar hur särskiljande ett ord är för ett visst dokument, dvs en god indikator för textens ämne

Utvärdering

- Två olika mått
  - ***Compression Ratio***: CR = (längd S) / (längd T)
  - ***Retention Ratio***: RR = (info i S) / (info i T)
- Mäta längd
  - Antal tecken? Antal ord?
- Mäta information
  - *Shannon Game*: återskapa ursprungstexten givet ett visst antal tecken
  - *Classification Game*: jämför klassifikation av sammanfattningar och originaltext
  - *Question Game*: besvara frågor om originaltexten givet sammanfattningar av olika längd

ROUGE-eval

- ROUGE (Recall Oriented Understudy for Gist Evaluation) är ett av de mest använda utvärderingspaketen för textsammanfattning
- **Mäter olika typer av överlapp mot en eller (oftast) flera referenssammanfattningar** (producerade av människor)
- ROUGE använder en rad olika mått
  - *ROUGE-N* mäter överlapp av *n*-gram
  - *ROUGE-S* mäter samförekomst av bigram
  - *ROUGE-SU* mäter samförekomst av bigram och unigram
  - *ROUGE-L* mäter antal längsta gemensamma delsträngar
  - *ROUGE-W* mäter antal längsta gemensamma delsträngar men premierar intilliggande delsträngar
- Kan efter registrering laddas ner från
  - http://www.berouge.com/Pages/default.aspx

SweSum I

- Sammanfattar nyhetstext och kortare rapporttexter på svenska, engelska, danska, norska, franska, tyska, spanska, italienska, grekiska och farsi online
- Använder de flesta av tidigare nämnda metoder
- Även formatering: utnyttjar HTML-taggar för
  - Fetstil
  - Nytt stycke
  - Styckerubriker
  - Dokumentrubrik

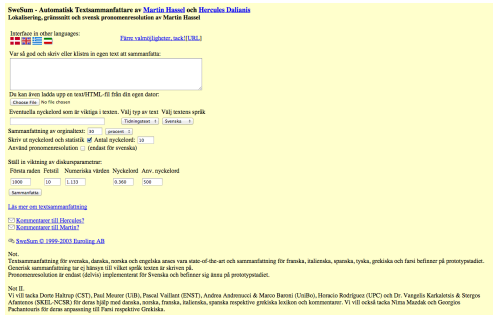
SweSum II

- Användaranpassade sammanfattningar (”slanting”)
  - Viktar upp **nyckelord** givna av användaren
- Naiv **kombinationsfunktion**
  - Använder både generella och genrespecifika indikatorer
  - Varje indikator viktas
  - Varje mening originaltexten får en rank baserat på indikatorerna
  - Högst rankade meningar extraheras och konkateneras i ursprunglig ordning till en sammanfattning

SweSum III

- För svenska
  - Lexikon med 700 000 innehållsord i böjd form, kopplade till lemmaform i ett lexikon
- Utvärdering
  - **Question Game**: 70-80% av frågor kring centrala faktum i en uppsättning texter kunde besvaras när 30% av originaltexten presenterades
- Implementering
  - Perl-CGI

SweSum UI



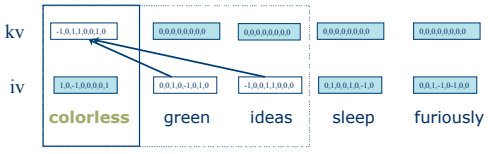
HolSum

- (Så gott som) **språkoberoende** textsammanfattare
  - ”små” språk saknar ofta stora mängder annoterad text, strukturerad data eller ”djupa” analysverktyg
- Avser producera **översiktssammanfattningar**
  - Försöker hitta en sammanfattning av en given längd som så lik originaltexten som möjligt
  - Med likhet avses här att **täcka samma ämnen i samma utsträckning som i originaltexten**

Att fånga orbetydelser

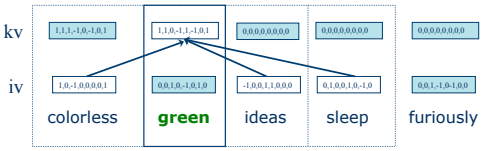
- HolSum använder **Random Indexing**, en teknik i familjen **distributionell lexikal semantik**, för att modellera innehållet i text och sammanfattning(ar)
- Samförekomststatistik
  - Ett **ordfönster** glider över texten och registrerar för varje ord vilka ord som förekommer tillsammans med det aktuella ordet
  - För varje observation av ett givet ord blir modellen av det ordets semantiska roll i språket mer statistisk välgrundad

Exempel: 2x2 Sliding Window



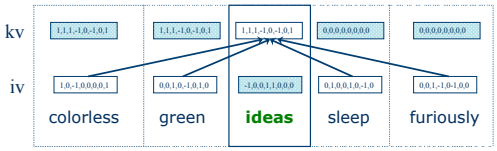
- **kv = kontextvektor**
  - Enbart nollor från början, uppdateras löpande
- **iv = indexvektor**
  - statisk, genereras vid första observationen

Kontextfönstret flyttar fram ett ord



- Vid varje observation adderas indexvektorn för de kringliggande orden till ordets kontextvektor
- På så vis innehåller ett ords kontextvektor spår av alla ord det förekommit tillsammans med

Fönstret flyttar ett ord till...



- Ord som förekommit i **snarlika kontexter** kommer att få **snarlika kontextvektorer**
- På så vis kan vi se om texter talar om samma sak även om de inte använder exakt samma ord

Att fånga textinnehåll

? Hur går vi då från att ha en konceptuell beskrivning av varje ord i texten till att få en beskrivning av textens innehåll – textens ämne

! Genom att summera de *tf*idf*-viktade kontextvektorerna för den aktuella textens alla ord

- Vi har nu en **innehållsvektor** för texten som representerar dess semantiska innehåll och som kan jämföras för likhet (**cosine**) med t.ex. en sammanfattning

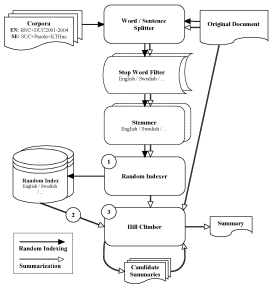
Att hitta en bättre sammanfattning

- Hill-climbing med början i **lead**
  Skapa en innehållsvektor för originaltexten och en för startsammanfattningen

  1. Förändra sammanfattningen genom att (slumpmässigt) ta bort en mening och lägga till en eller flera meningar ur originaltexten
  2. Jämför den nya sammanfattningens innehållsvektor med originaltextens
  3. Behåll den bästa kandidaten (gamla eller nya)

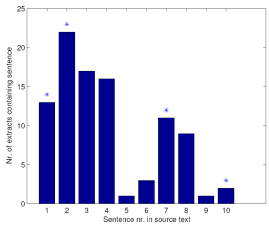
  Upprepa 1-3 tills ingen bättre sammanfattning hittas (dvs ett **lokalt optimum** uppnåtts)

HolSum: Systemarkitektur



1. Semantisk modellering
2. Tillämpning av semantisk modell
3. Semantisk navigering genom rymden av möjliga sammanfattningar

HolSum: Utvärdering



- Antal extrakt producerade av människor som inkluderar respektive mening ur en text från den svenska korpusen
- Meningar markerade med * är de som HolSum valde
- Det finns totalt 27 extrakt producerade för denna text

Utmaningar

- En av de stora utmaningarna med extraktionsbaserad textsammanfattning är att få texten sammanhängande – dvs **koherent**
- En text har **koherens** när innehållet bildar en sammanhängande helhet
- En text har **kohesion** när det finns explicita kopplingar mellan olika led, t.ex. genom
  - Sambandsled (= ”anaforiska” led som syftar tillbaka på något givet eller känt)
  - Satskonnektiver (de fem vanligaste är **icke**, **och**, **eller**, **om… så…** och **om och endast om**)

Koreferens

- Referentbestämning
  - Att bestämma vad (eller vem) ett givet uttryck relaterar till
- Koreferens
  - Två eller fler refererande uttryck relaterar till samma referent

Pelle var snabb. **Han** sprang snabbast i klassen.
antecedent — anaforiskt uttryck

Varför är detta ett problem?

- Hängande anaforer (tillbaks till Pelle)

  Pelle var snabb. **Han** sprang snabbast i klassen.

- Vad händer om vår ATS väljer att enbart inkludera den andra meningen?

  ~~**Pelle** var snabb.~~ **Han** sprang snabbast i klassen.

- Hur ska vi nu veta vem **Han** refererar till?

SweSum: Utan referentbestämning

Analysera mera!
Regi: Harold Ramis
Medv: Robert De Niro, Billy Crystal, Lisa Kudrow
Längd: 1 tim, 45 min
…
Ett av många skäl att glädjas åt Analysera mera är att Robert De Niro här verkligen utövar skådespelarkonst igen. **Han** accelererar emotionellt från 0 till 100 på ingen tid alls, för att sedan kattmjukt bromsa in och parkera, lugnt och behärskat. Och **han** är tämligen oemotståndlig. Här har **han** åstadkommit ännu en intelligent komedi för alla oss vänner av intelligens och komedi, gärna i kombination.
SvD 99-10-08

SweSum: Med referentbestämning

Analysera mera!
Regi: Harold Ramis
Medv: Robert De Niro, Billy Crystal, Lisa Kudrow
Längd: 1 tim, 45 min
…
Ett av många skäl att glädjas åt Analysera mera är att Robert De Niro här verkligen utövar skådespelarkonst igen. **Robert** accelererar emotionellt från 0 till 100 på ingen tid alls, för att sedan kattmjukt bromsa in och parkera, lugnt och behärskat. Och **Robert** är tämligen oemotståndlig. Här har **Harold** åstadkommit ännu en intelligent komedi för alla oss vänner av intelligens och komedi, gärna i kombination.
SvD 99-10-08

Typer av refererande uttryck

Fem vanliga typer av referande uttryck	
Typ	*Exempel*
Indefinita nominalfraser	Jag såg **en** Ford idag.
Definita nominalfraser	Jag såg en Ford idag. **Den** forden var vit.
Pronomen	Jag såg en Ford idag. **Den** var vit.
Demonstrativer	Jag tycker bättre om **denna**.
One-anaphora	Jag såg 6 Ford-bilar idag. Jag vill också ha **en**.
Tre typer av refererande uttryck som försvårar referentbestämning	
Typ	*Exempel*
Funktionella (“inferrables”)	Jag köpte nästan en Ford, men **dörren** var bucklig.
Diskontinuerliga	Peter och Paul har varsin Ford. **De** kör **dem** ofta.
Generiska	Jag såg 6 Ford-bilar idag. **De** är häftiga bilar.

Villkor för anaforisk koreferens

- Morfologiska egenskaper
  - Antecedent och anafor ska ha samma **genus** och **numerus**

    Bollarna träffade **glaset**. **Det** gick sönder.
- Syntaktiska egenskaper
  - Antecedentens och anaforens **syntaktiska** positioner måste tillåta koreferens

    Peter gav honom en present → Peter ≠ honom
- Koreferens kan kräva betydande **omvärldskunskap**

  Stefan Löfven får en knepig start. Den nya **statsministern** måste kohandla med båda blocken.

Mitkovs Limited Knowledge Approach

- Kräver inte parsning, utan endast ordklasstaggning och nominalfraschunkning
- Mer intuitivt viktningssystem än en del andra metoder (t.ex. Lappin & Leass, 1994)
- Missar dock grammatiska roller (vilket t.ex Lappin & Leass hanterar)

- Har implementerats och utvärderats för engelska, polska och arabiska bruksanvisningar
- Begränsad utvärdering för svenska (Algotsson, 2007)

Mitkovs algoritm (1997)

1. Ta ordklasstaggad text som indata
2. Identifiera nominalfraser i som mest två meningar bort från den aktuella anaforen
3. Kontrollera kongruens i genus och numerus
4. Tillämpa genrespecifika antecedentindikatorer
5. Välj som antecedent den kandidat som har högstpoäng på indikatorerna
6. Tillämpa **tie-break** om flera kandidater har samma poäng

Mitkovs Antecedentindikatorer I

- *First Noun Phrase*: +1 till första NP i mening
- *Indicating verbs*: +1 till NP som direkt följer på ett verb i en fördefinierad lista (domänspecifikt)
- *Term Preference*: +1 till NP som tillhör textens genre
- *Section Heading Preference*: +1 till NP som förekommer i rubriken till det stycke aktuellt pronomen förekommer i
- *Lexical Reiteration*: +2 till NP som som förekommer två eller fler gånger i samma stycke som aktuellt pronomen, +1 till de som förekommer en gång

Mitkovs Antecedentindikatorer I

- *Immediate Reference*: +2 till NP som förekommer precis före aktuellt pronomen, *i samma mening*
- *Collocation Match*: +2 till NP som har samma kollokationsmönster som aktuellt pronomen

- Straffpoäng
  - *Indefiniteness*: -1 till indefinita NP
  - *Prepositional Noun Phrases*: -1 NP i prepositionsfraser

Tie-break i Mitkovs algoritm

- Om två eller fler nominalfraser hamnar på samma högsta poäng, föredra då den kandidat som
  - Har högst *immediate reference score*
  - Har högst *collocation pattern score*
  - Har högst *indicating verb score*
  - Ligger närmast före anaforen i fråga

När behövs referentbestämning?

Stockholms universitet

- Som vi redan har sett är referentbestämning viktigt vid extraktionsbaserad textsammanfattning
- Men referentbestämning behövs även vid
  - Maskinöversättning
  - Informationsextraktion
  - Och i nästan alla tillämpningar av Natural Language Understanding (ett underområde till AI)

Referentbestämning – Verktyg och korpusar

Stockholms universitet

- Engelska
  - GATE (NLP-toolkit skrivet i Java) https://gate.ac.uk/
  - Java-RAP http://aye.comp.nus.edu.sg/~qiu/NLPTools/JavaRAP.html
  - GUITAR http://sourceforge.net/projects/guitar-essex/
- Svenska
  - SUC-CORE, en delmängd av Stockholm-Umeå korpus 2.0 (inkluderad även i 3.0)

Textsammanfattning – Verktyg

Stockholms universitet

- Engelska
  - AutoSummarize (i Microsoft Word)
  - SUMMA (plug-in till GATE) http://www.taln.upf.edu/pages/summa.upf/
  - MEAD http://www.summarization.com/mead/
  - Open Text Summarizer http://libots.sourceforge.net/
  - Columbia Newsblaster http://newsblaster.cs.columbia.edu/
- Svenska
  - SweSum http://swesum.nada.kth.se/

Textsammanfattning – Korpusar

Stockholms universitet

- Engelska
  - DUC corpus http://duc.nist.gov
  - SummBank corpus http://www.summarization.com/summbank
  - SUMMAC corpus Skicka förfrågan till mani@mitre.org
  - <Text+Abstract+Extract> corpus Skicka förfrågan till marcu@isi.edu
- Svenska (ej allmänt tillgängliga)
  - KTHnc-QA (Dalianis & Hassel, 2001): Delmängd av KTH News Corpus (Hassel, 2001) med tillhörande frågor och svar gällande centrala fakta i texterna
  - KTHxc (Hassel & Dalianis, 2005): KTH eXtract Corpus, texter med tillhörande extrakt producerade av människor