KTH Data- och systemvetenskap

Danish extract corpus

Hercules Dalianis

Department of Computer
and System Sciences

Stockholm University and KTH

Email:hercules@kth.se

KTH Data- och systemvetenskap

DanSum

- DanSum - text summarizer ported from SweSum
  - Sept 2002
- DanSum summarizes Danish news text
- DanSum requires evaluation
- Evaluation corpus needed
- Berlingske Tidende

DanSum - automatisk resummering af danske tekster

Sådan bruger du DanSum

Skriv adressen på en hjemmeside (et dokument) og klik derefter på "**Resumer**":

http://www.cst.dk/defsum/Information.html

Skriv eventuelle nøgleord der skal ledes efter i teksten.

Vælg teksttype

Avistekst

Resummering af originaltekst: 34 %

Udskriv nøgleord og statistik ☑

Resumer

Vejen til flere valgmuligheder

Bemærk:
DanSum er et resultat af et samarbejde mellem Center for Sprogteknologi (CST) i København og Kungliga Tekniska Høgskolan (KTH) i Stockholm.
CST's arbejde er udført inden for projektet DefSum der er støttet af Danmarks Elektroniske Forskningsbibliotek (DEF).

Kontakt Hercules Dalianis, KTH?
Kontakt Martin Hassel, KTH?
Kontakt Juergen Wedekind, CST?
Kontakt Dorte Haltrup, CST?

KTH Data- och systemvetenskap

Collecting extract corpus

- KTH extract tool
- 10 news text from Berlingske Tidende
- Experiment december 2004
- 15 informants mostly CST personnel
- We obtained 135 Danish extract
- Average length of 32 percent of original texts
- 67 percent overlap at average length

KTH extraktkorpus

Nedan visas tre stycken selektionsmenyer. Den första för *språk* tillgängliga i korpusen, den andra för *texttyper* tillgängliga för ett valt språk och den tredje för *texter* tillgängliga för en viss texttyp (för ett visst språk). Du kan använda dessa för att orientera dig korpusen och välja ut specifika filer som du vill titta på.

danska | nyhedstekster | text001.htm (15)

Filnamn (text)	Antal extrakt	Kortaste extrakt	Längsta extrakt	Meddellängd	Överlapp alla röster	Överlapp, medellängd	Abstrakt	Ändrad
text001.htm	15	14%	42%	34%	44%	76%		
text002.htm	14	18%	53%	35%	40%	66%		
text003.htm	15	15%	40%	32%	34%	67%		
text004.htm	12	10%	34%	23%	32%	55%		
text005.htm	13	9%	53%	30%	39%	60%		
text006.htm	12	15%	46%	28%	37%	67%		
text007.htm	11	15%	51%	33%	42%	63%		
text008.htm	14	11%	59%	37%	31%	79%		
text009.htm	15	24%	52%	38%	37%	73%		
text010.htm	14	19%	45%	31%	38%	65%		
Totalt/Medel	135	15%	48%	32%	38%	67%		

KTH Data- och systemvetenskap

Nedan visas 25% av originaltexten. Denna sammanfattning representerar det bästa extraktet i enlighet med majoritetsbeslut baserat på 12 extrakt. Värdet inom hakparenteser före varje mening representerar antalet gånger denna mening blivit utvald till ett extrakt.

Denna texts ID är: danska->nyhedstekster->text004.htm

Ideal:

1[11] Flyvevåbnets F16-fly er først om et par år bragt teknologisk helt på højde med de tilsvarende amerikanske jagere .
2[6] Såfremt NATO beder danske F16 deltage i angreb mod mål på jorden i Serbien og Kosovo , vil de danske fly kunne aflevere en række forskellige bomber og missiler .
3[7] Præcisionsbombning med laserstyrede bomber kan kun ske med hjælp fra andre fly eller med hjælp fra soldater - forward air controllers - på jorden .
7[8] Flyvevåbnet har bestilt et antal laserpods , men de bliver først klar til anvendelse i år 2001 .
9[6] Men vi opfylder fuldt ud NATOs standarder , og har samme udstyr som andre europæiske NATO-landes flyvevåben .
14[8] Voldsomme våben Også de danske flys elektroniske udstyr , blandt andet dets cockpit , halter teknologisk lidt bagud i sammenligning med amerikanske fly .
17[6] Men F16-flyene , der efter de foreliggende oplysninger fra Forsvarskommandoen hidtil alene har udført defensive opgaver , råder trods manglerne over voldsomme våben .
18[7] Udover at kaste bomber - styrede eller ustyrede - medbringer F16-flyene efter Berlingske Tidendes oplysninger det topmoderne Maverick-missil til Italien .
28[5] Et nok så frygteligt våben er CRV-7 raketterne .
30[5] F16 kan medbringe 76 stykker , fordelt i fire beholdere , og de kan affyres enkeltvis , i grupper eller samtidigt .
33[5] Endelig råder flyvevåbnet over klyngebomber .
44[5] årsagen til NATOs anmodning om flere fly til operationerne mod det serbiske militær er ifølge kaptajn Jesper Myrthue , dels at de mange operationer slider på både maskiner og piloter , og dels at NATO med flere fly kan sende flere angrebsbølger med kortere varslingstid mod mål i Kosovo og Serbien .

Visa sammanfattning på 25 procent (268 ord av 1057).
Sammanfattningen ovan är baserad på 12 extrakt.
Kortaste extraktet representerat ovan är 10%, längsta är 34% och medellängden är 23%.
Täckningen för Ideal sammanfattning (majoritetsval) är 51% och precisionen för densamma är 55%.
Täckningen för Baseline 1 (slumpvis distribution) är 26% och precisionen för densamma är 28%.
Täckningen för Baseline 2 (inledande meningar, text) är 40% och precisionen för densamma är 43%.
Täckningen för Baseline 3 (inledande meningar, stycke) är 40% och precisionen för densamma är 43%.

KTH Data- och systemvetenskap

DanSum and extract corpus

- Comparing performance of DanSum with Danish extract corpus
- Majority vote extract at average length ideal extract-Gold Standard
- DanSum - with input of original text and average length parameters

KTH Data- och systemvetenskap

Quality

Correlation

Danish Extracts Filename	No of extracts	Average extract length	Overlap of all votes	Overlap at average length	Comparison with DanSum summary: Overlap at sentence level	word level	word frequency	truncated word level	truncated word frequency
text001.htm	15	34%	44%	76%	57%	77%	67%	78%	67%
text002.htm	14	35%	40%	66%	67%	56%	47%	59%	48%
text003.htm	15	32%	34%	67%	18%	56%	43%	56%	42%
text004.htm	12	23%	32%	55%	44%	55%	44%	61%	48%
text005.htm	13	30%	39%	60%	57%	54%	43%	56%	45%
text006.htm	12	28%	37%	67%	60%	68%	57%	69%	56%
text007.htm	11	33%	42%	63%	67%	69%	63%	71%	65%
text008.htm	14	37%	31%	79%	40%	46%	31%	48%	30%
text009.htm	15	38%	37%	73%	67%	81%	71%	81%	70%
text010.htm	14	31%	38%	65%	40%	73%	65%	74%	64%
Total/Average	135	32%	37%	67%	52%	64%	53%	65%	54%

- Problems with tokenization i Danish extract corpus.
- One extract unit contain many sentences
- In text 1 does extraction unit 4, 6 & 11 have two sentences each (6 & 11 with colon :),

KTH Data- och systemvetenskap

KTH extraktkorpus

Nedan visas 34% av originaltexten. Denna sammanfattning representerar det bästa extraktet i enlighet med majoritetsbeslut baserat på 15 extrakt. Värdet inom hakparenteser före varje mening representerar antalet gånger denna mening blivit utvald till ett extrakt.

Denna texts ID är: danska->nyhedstekster->text001.htm

Ideal:

0[13] **Udnævnelse af EU-Kommission trækker ud**
1[15] BRUXELLES De 15 regeringer i Den Europæiske Union håber efter onsdag aftens topmøde i Bruxelles , at en ny beslutningsdygtig EU-Kommission kan være på plads til august .
3[10] De 15 landes stats- og regeringschefer mødtes i går første gang med deres kandidat til formandsposten , italieneren Romano Prodi , som for tre uger siden blev nomineret til jobbet i al hast på et topmøde i Berlin .
4[12] Alt peger på , at Europa-Parlamentet på sin næste samling i maj vil godkende Prodi som formand for Kommissionen , og derefter skal han i samarbejde med de enkelte regeringer i gang med at finde de 19 menige kommissærer til fremtidens Kommission . ™ Jeg håber , at Europa-Parlamentet kan begynde høringerne af de enkelte kommissærer i juli , så parlamanterikerne kan godkende den samlede Kommission i første uge af august , ¥ sagde den tyske kansler og nuværende EU-formand Gerhard Schrñder efter mødet .
7[12] Santer måneder endnu En række parlamentarikere er stærkt utilfredse med , at den nuværende skandaliserede Kommission risikerer at skulle sidde i endnu fire-fem måneder - og især den fungerende formand Jacques Santers og den franske kommissær Edith Cressons tilstedeværelse er stærkt uønsket .
14[10] Dansk kommissær Statsministeren ville fortsat ikke antyde , i hvilken retning hans egne overvejelser om en ny dansk kommissær går .
18[8] Inspirationspapir Nyrup præsenterede på topmødet sine kolleger og Prodi for et såkaldt inspirationspapir , som handler om en hovedrengøring i Kommissionen .

Visa sammanfattning på 34 procent (243 ord av 709).
Sammanfattningen ovan är baserad på 15 extrakt.
Kortaste extraktet representerat ovan är 14%, längsta är 42% och medellängden är 34%.
Täckningen för Ideal sammanfattning (majoritetsval) är 64% och precisionen för densamma är 76%.
Täckningen för Baseline 1 (slumpvis distribution) är 29% och precisionen för densamma är 34%.
Täckningen för Baseline 2 (inledande meningar, text) är 53% och precisionen för densamma är 63%.
Täckningen för Baseline 3 (inledande meningar, stycke) är 53% och precisionen för densamma är 63%.

KTH Data- och systemvetenskap

KTH extraktkorpus

Nedan visas 34% av originaltexten. Denna sammanfattning representerar det bästa extraktet i enlighet med majoritetsbeslut baserat på 15 extrakt. Värdet inom hakparenteser före varje mening representerar antalet gånger denna mening blivit utvald till ett extrakt.

Denna texts ID är: danska->nyhedstekster->text001.htm

Ideal:

0[13] **Udnævnelse af EU-Kommission trækker ud**
1[15] BRUXELLES De 15 regeringer i Den Europæiske Union håber efter onsdag aftens topmøde i Bruxelles , at en ny beslutningsdygtig EU-Kommission kan være på plads til august .
3[10] De 15 landes stats- og regeringschefer mødtes i går første gang med deres kandidat til formandsposten , italieneren Romano Prodi , som for tre uger siden blev nomineret til jobbet i al hast på et topmøde i Berlin .
4[12] Alt peger på , at Europa-Parlamentet på sin næste samling i maj vil godkende Prodi som formand for Kommissionen , og derefter skal han i samarbejde med de enkelte regeringer i gang med at finde de 19 menige kommissærer til fremtidens Kommission . ™ Jeg håber , at Europa-Parlamentet kan begynde høringerne af de enkelte kommissærer i juli , så parlamanterikerne kan godkende den samlede Kommission i første uge af august , ¥ sagde den tyske kansler og nuværende EU-formand Gerhard Schrñder efter mødet .
7[12] Santer måneder endnu En række parlamentarikere er stærkt utilfredse med , at den nuværende skandaliserede Kommission risikerer at skulle sidde i endnu fire-fem måneder - og især den fungerende formand Jacques Santers og den franske kommissær Edith Cressons tilstedeværelse er stærkt uønsket .
14[10] Dansk kommissær Statsministeren ville fortsat ikke antyde , i hvilken retning hans egne overvejelser om en ny dansk kommissær går .
18[8] Inspirationspapir Nyrup præsenterede på topmødet sine kolleger og Prodi for et såkaldt inspirationspapir , som handler om en hovedrengøring i Kommissionen .

Visa sammanfattning på 34 procent (243 ord av 709).
Sammanfattningen ovan är baserad på 15 extrakt.
Kortaste extraktet representerat ovan är 14%, längsta är 42% och medellängden är 34%.
Täckningen för Ideal sammanfattning (majoritetsval) är 64% och precisionen för densamma är 76%.
Täckningen för Baseline 1 (slumpvis distribution) är 29% och precisionen för densamma är 34%.
Täckningen för Baseline 2 (inledande meningar, text) är 53% och precisionen för densamma är 63%.
Täckningen för Baseline 3 (inledande meningar, stycke) är 53% och precisionen för densamma är 63%.

KTH Data- och systemvetenskap

DanSum performance

- Around 50 percent sentence overlap with majority votes
- Informants agree to 67 percent