Meningsåterställare via n-gramstatistik

Språkteknologi DD2418 HT2016

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Problem och metod

Språkteknologi DD2418 HT2016
Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Problem och metod

- Problem: mening som tappat ordning mellan ord

  "dags minut senare en igen det var"

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Problem och metod

- Problem: mening som tappat ordning mellan ord

  "dags minut senare en igen det var"

- Metod: n-gramstatistik över svenska, testa alla permutationer

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Problem och metod

- Problem: mening som tappat ordning mellan ord

  "dags minut senare en igen det var"

- Metod: n-gramstatistik över svenska, testa alla permutationer

  "en stor hund", "en hund stor", "stor en hund", ...

```
ngramscore("en stor") + ngramscore("stor hund")
```


Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Källa/källor till statistik

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Källa/källor till statistik

- Språkbanken (Göteborgs Universitet)

  `https://spraakbanken.gu.se`

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Källa/källor till statistik

- Språkbanken (Göteborgs Universitet)

  `https://spraakbanken.gu.se`

- Korpusar i gigantiska XML-filer

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Källa/källor till statistik

- Språkbanken (Göteborgs Universitet)

  `https://spraakbanken.gu.se`

- Korpusar i gigantiska XML-filer

- Inga verktyg: egna verktyg! ... Dokumentation?

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Korpusparsning

Språkteknologi DD2418 HT2016
Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Korpusparsning

- Python - minnesläcka i biblioteket? (lxml)

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Korpusparsning

- Python - minnesläcka i biblioteket? (lxml)

- C + glib + libxml2

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Korpusparsning

- Python - minnesläcka i biblioteket? (lxml)
- C + glib + libxml2
- mkfifo + bzcat!

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Korpusparsning

- Python - minnesläcka i biblioteket? (lxml)
- C + glib + libxml2
- mkfifo + bzcat!

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Korpusparsning

- Python - minnesläcka i biblioteket? (lxml)
- C + glib + libxml2
- mkfifo + bzcat!

/home/lur/Desktop/sprakproj/sbc-txt/sbc_triends.txt_proc.txt - SciTE

File Edit Search View Tools Options Language Buffers Help

sbc_triends.txt_proc.txt

```
9148 sa galne hugo 1
9149 sa glada molnet 2
9150 sa greven torrt 1
9151 sa han allvarligt 1
9152 sa han argt 1
9153 sa han bara 4
9154 sa han bestämt 1
9155 sa han det 1
9156 sa han egentligen 1
9157 sa han förtvivlat 1
9158 sa han glatt 2
9159 sa han hest 1
9160 sa han högt 1
9161 sa han igen 1
9162 sa han ilsket 1
9163 sa han imponerat 1
9164 sa han kom 1
9165 sa han lågt 1
9166 sa han nervöst 1
9167 sa han oroligt 1
9168 sa han sedan 1
```

[INS] /home/lur/Desktop/sprakproj/sbc-txt/sbc_triends.txt_proc.txt @ 1, 1 (0)

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Algoritm

Språkteknologi DD2418 HT2016
Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Algoritm: uppslag i statistikfilerna

- Binärsökning!

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Algoritm: steg 1

- `for x in itertools.permutations(scrambledtext.split(' '))`

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Algoritm: steg 1

- `for x in itertools.permutations(scrambledtext.split(' '))`

- `bigrams = zip(x, x[1:])`
  `... [['jag', 'har'], ['har', 'en'], ['en', 'snäll'], ...`

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Algoritm: steg 1

- `for x in itertools.permutations(scrambledtext.split(' '))`

- `bigrams = zip(x, x[1:])`
  `... [['jag', 'har'], ['har', 'en'], ['en', 'snäll'], ...`

- `trigrams = zip(x, x[1:], x[2:])`
  `... [['jag', 'har', 'en'], ['har', 'en', 'snäll'], ...`

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Algoritm: steg 1

- `for x in itertools.permutations(scrambledtext.split(' '))`

- `bigrams = zip(x, x[1:])`
  `... [['jag', 'har'], ['har', 'en'], ['en', 'snäll'], ...`

- `trigrams = zip(x, x[1:], x[2:])`
  `... [['jag', 'har', 'en'], ['har', 'en', 'snäll'], ...`

- Steg 1: finn delmängd permutationer med lägst antal "nollor"

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Algoritm: steg 2

- Steg 2: välj den "bästa" permutationen funnen i steg 1

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Algoritm: steg 2

- Steg 2: välj den "bästa" permutationen funnen i steg 1

- Summan av frekvenserna för vardera (bi|tri)gram

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Algoritm: steg 2

- Steg 2: välj den "bästa" permutationen funnen i steg 1
- Summan av frekvenserna för vardera (bi|tri)gram
- Ökad vikt för trigram, ökad vikt för (bi|tri)gram som inleder eller avslutar mening

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Algoritm: steg 2

- Steg 2: välj den "bästa" permutationen funnen i steg 1
- Summan av frekvenserna för vardera (bi|tri)gram
- Ökad vikt för trigram, ökad vikt för (bi|tri)gram som inleder eller avslutar mening
- Demo!

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Algoritm: steg 2

- Steg 2: välj den "bästa" permutationen funnen i steg 1
- Summan av frekvenserna för vardera (bi|tri)gram
- Ökad vikt för trigram, ökad vikt för (bi|tri)gram som inleder eller avslutar mening
- Bra metod? Utvärdering

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Utvärdering

Språkteknologi DD2418 HT2016
Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Utvärdering

- Ett kvalitetsmått

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Utvärdering

- Ett kvalitetsmått

```
correct = "sverige gjorde en bra match"
ans     = "sverige gjorde en match bra"
```


Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Utvärdering

- Ett kvalitetsmått

```
correct = "sverige gjorde en bra match"
ans     = "sverige gjorde en match bra"

ans_substr = ["sverige gjorde", "sverige gjorde en",
              "sverige gjorde en match", "sverige gjorde en
               match bra", "gjorde en", "gjorde en match" ...
```


Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Utvärdering

- Ett kvalitetsmått

```
correct = "sverige gjorde en bra match"
ans     = "sverige gjorde en match bra"

ans_substr = ["sverige gjorde", "sverige gjorde en",
              "sverige gjorde en match", "sverige gjorde en
               match bra", "gjorde en", "gjorde en match" ...

score = num_match(correct, ans_substr) / num_substr(correct)
```


Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Utvärdering

- Testmeningar från källa som inte använts för statistiken

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Utvärdering

- Testmeningar från källa som inte använts för statistiken

- Gruppera testmeningarna efter antal ord (3 ... 8)

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Utvärdering

- Testmeningar från källa som inte använts för statistiken
- Gruppera testmeningarna efter antal ord (3 ... 8)
- Jämför med att slumpa ordningen (Python3 random.shuffle, Linux) - nollhypotes?

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Utvärdering

- Testmeningar från Språkbankens korpus "LäSBarT - Lättläst svenska och barnbokstext"

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Utvärdering

- Testmeningar från Språkbankens korpus "LäSBarT - Lättläst svenska och barnbokstext"

Antal ord	Antal testmeningar
3	1561
4	1627
5	1549
6	1279
7	1028
8	720

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Utvärdering

- Testmeningar från Språkbankens korpus "LäSBarT - Lättläst svenska och barnbokstext"

Antal ord	Antal testmeningar
3	1561
4	1627
5	1549
6	1279
7	1028
8	720

```
de lyckades till sist
de verkar trivas tillsammans
det skriver tv4s webbnyheter
flera andra blev skadade
han kan knappt andas
har mormor blivit tokig
ingen svensk var ombord
men det finns problem
men folket sade nej
men polisen stoppade henne
men striderna har fortsatt
precis som i brasilien
riksdagen styr hela sverige
sitter i sin rullstol
vad kul det var
...
```


Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Resultat

- Statistik baserat på korpusen “GP2010”

Antal ord i mening	Descrambler	Random
3	0.545	0.294
4	0.507	0.173
5	0.437	0.105
6	0.372	0.072
7	0.303	0.056
8	0.255	0.039

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Resultat

- Statistik baserat på korpusen “Svenska Wikipedia”

Antal ord i mening	Descrambler	Random
3	0.464	0.296
4	0.444	0.167
5	0.395	0.100
6	0.323	0.070
7	0.273	0.051
8	0.237	0.039

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Resultat

- Statistik baserat på stor sammanslagen* korpus

Antal ord i mening	Descrambler	Random
3	0.629	0.290
4	0.540	0.157
5	0.457	0.104
6	0.374	0.074
7	0.308	0.050
8	0.266	0.042

*bloggmix2011 + gp2010 + gp2011 + svwikipedia

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Resultat

- Statistik baserat på korpusen “LäSBarT” (oops!)

Antal ord i mening	Descrambler	Random
3	0.985	0.270
4	0.986	0.172
5	0.979	0.103
6	0.989	0.067
7	0.980	0.051
8	0.987	0.042

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Slutsats

Språkteknologi DD2418 HT2016
Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Slutsats

- Algoritmen presterar bättre än random.shuffle

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Slutsats

- Algoritmen presterar bättre än random.shuffle

- Stor skillnad i resultat för olika statistikbaser och olika språkstilar för indata

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Slutsats

- Algoritmen presterar bättre än random.shuffle
- Stor skillnad i resultat för olika statistikbaser och olika språkstilar för indata
- Algoritmen presterar bättre på kortare meningar (eller?)

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Slutsats

- Algoritmen presterar bättre än random.shuffle
- Stor skillnad i resultat för olika statistikbaser och olika språkstilar för indata
- Algoritmen presterar bättre på kortare meningar (eller?)
- Det finns massor av “oavgörbara” meningar – svårt utvärdera

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Diskussion

Språkteknologi DD2418 HT2016
Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Diskussion

- Kvalitetsmåttet är strängt

```
Correct:  “det har ingen annan klarat”
Answer:   “ingen annan har klarat det”

Score = 0.10
```


Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Diskussion

- Kvalitetsmåttet är strängt

```
Correct:  “det har ingen annan klarat”
Answer:   “ingen annan har klarat det”

Score = 0.10
```

- Borde testa fler uppsättningar testmeningar, testmeningar i olika kategorier (nyhetstext, prosa, ...)

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Diskussion

- Kvalitetsmåttet är strängt

```
Correct:  “det har ingen annan klarat”
Answer:   “ingen annan har klarat det”

Score = 0.10
```

- Borde testa fler uppsättningar testmeningar, testmeningar i olika kategorier (nyhetstext, prosa, ...)

- Prova att inkludera 4-gram?

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Vad har vi lärt oss om språkteknologi?

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Vad har vi lärt oss om språkteknologi?

- Stora filer, stora datamängder, inte lägga på git!

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Vad har vi lärt oss om språkteknologi?

- Stora filer, stora datamängder, inte lägga på git!

- Nedladdning tar tid, parsning tar tid, processande tar tid

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Vad har vi lärt oss om språkteknologi?

- Stora filer, stora datamängder, inte lägga på git!
- Nedladdning tar tid, parsning tar tid, processande tar tid
- Statistisk metod kräver stort antal tester - tar tid

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Vad har vi lärt oss om språkteknologi?

- Stora filer, stora datamängder, inte lägga på git!
- Nedladdning tar tid, parsning tar tid, processande tar tid
- Statistisk metod kräver stort antal tester - tar tid
- Svårt att utvärdera när det kan finnas flera “rätt”

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>

Meningsåterställare via n-gramstatistik

Frågor?

Någon som vill testa en mening?

Mikael Forsberg <miforsb@kth.se>
Robin Gunning <rgunning@kth.se>