T-(538|725)-MALV, Málvinnsla
Talningar og N-stæður

Hrafn Loftsson[1] Hannes Högni Vilhjálmsson[1]

[1]Tölvunarfræðideild, Háskólinn í Reykjavík

September 2007

Outline

Outline

Talningar og orðarunur

Orðastæður (e. collocations)

- Sambönd (eða röð) orða sem mynda merkingarlega heild og koma iðulega fyrir sem samstæða innan setningar.
- Röð orða sem birtast saman síendurtekið.
- Orðastæðurnar eru ekki tilviljanakenndar.
- Getur verið nauðsynlegt að finna orðastæður, t.d. í orðabókagerð.
- Dæmi: “crystal clear”, “cosmetic surgery”, “blonde hair”, “oft og tíðum”, “veikur hlekkur”.

Talningar og orðarunur

Mállíkan (e. language model)

- Tölfræðileg nálgun á tíðni og röð orða.
- Oft notað til að spá fyrir um næsta orð þegar undanfarandi röð orða er þekkt.
- Mjög mikið notað í ýmiss konar málvinnslu:
  - Talgreiningu, orðabókagerð, mörkun, setningagreiningu, merkingargreiningu, þýðingum, o.s.frv.

Orðmyndir og tókar

Word types (orðmyndir)

- Ólíkar orðmyndir.
- Í málheildinni *Íslensk orðtíðnibók* eru 59.358 orðmyndir.

Word tokens (tókar/lesmálsorð)

- Öll orð.
- Í málheildinni *Íslensk orðtíðnibók* eru 590.297 tókar.

Dæmi

- Þetta er skóli. Anna sá skóla. Atli sá skóla.
- 12 tókar, 7 orðmyndir.

Orðmyndir og tókar

Word types (orðmyndir)

- Ólíkar orðmyndir.
- Í málheildinni *Íslensk orðtíðnibók* eru 59.358 orðmyndir.

Word tokens (tókar/lesmálsorð)

- Öll orð.
- Í málheildinni *Íslensk orðtíðnibók* eru 590.297 tókar.

Dæmi

- Þetta er skóli. Anna sá skóla. Atli sá skóla.
- 12 tókar, 7 orðmyndir.

N-stæður

- Runur af N orðum (sem standa saman).
- Einstæður: Unigrams
- Tvístæður: Bigrams
- Þrístæður: Trigrams
- Fjórstæður: Fourgrams
- o.s.frv.

N-stæður

- Þetta er skóli. Anna sá skóla. Atli sá skóla.
- Tvístæður: “Þetta er”, “er skóli”, “skóli .”, “. Anna”, “Anna sá”, “sá skóla”, o.s.frv.
- Þrístæður: “Þetta er skóli”, “er skóli .”, “skóli . Anna”, “. Anna sá”, “Anna sá skóla”, “sá skóla .”, o.s.frv.

Outline

Fleiri Unix/Linux tól

sort

- Stafrófsröðun:
  - sort inputfile > outputfile
  - Getið þið fengið þetta til að raða íslenskum texta rétt?
  - T.d. í Cygwin eða Linux?
- Öfug röð:
  - sort -r inputfile > outputfile
- Númerísk röðun
  - sort -n inputfile > outputfile

Fleiri Unix/Linux tól

uniq

- Eyðir öllum nema einni línu úr röð eins lína.
- uniq inputfile > outputfile
- sort inntak.txt | uniq > uttak.txt
- Með tíðnitölu:
  - uniq -c inputfile > outputfile
- Taka saman tíðni orða
  - sort inntak.txt | uniq -c | sort -nr > uttak.txt

Gerð N-stæðulíkana

Einstæðulíkan

- Inntak: Málheild.
  1. Tilreiðing – eitt orð (tóki) í hverri línu.
  2. Talning.

Auðvelt í Unix/Linux

- G.r.f. að skrá *malheild.wrd* innihaldi einn tóka í línu.
- sort mailheild.wrd | uniq -c | sort -nr > malheild.freq

Gerð N-stæðulíkana

Einstæðulíkan

- Inntak: Málheild.
  1. Tilreiðing – eitt orð (tóki) í hverri línu.
  2. Talning.

Auðvelt í Unix/Linux

- G.r.f. að skrá *malheild.wrd* innihaldi einn tóka í línu.
- sort mailheild.wrd | uniq -c | sort -nr > malheild.freq

Einstæðulíkan í Perl (4.4.3 í kennslubók)

```
$text = <>;
while ($line = <>) { $text .= $line}
$text =~ tr /a-záðéíóýúæöþA-ZÁÐÉÍÓÝÚÆÖÞ0-9.,!?\-:;/\n/cs;   # ófullkomin
$text =~ s/([,.?!:;()\-])/\n$1\n/g;                           # tilreiðing
$text =~ s/\n+/\n/g;

@words = split(/\n/, $text);
for ($i=0; $i <= $#words; $i++) {
    if (!exists($frequency{$words[$i]})) {$frequency{$words[$i]} = 1;}
    else {$frequency{$words[$i]}++;}
}

foreach $word (sort keys %frequency) {
    print "$frequency{$word} $word\n";
}
```

Fleiri Unix/Linux tól

head og *tail*

- head -3 < inntak.txt
  - Skilar fyrstu þremur línunum.
- tail -2 < inntak.txt
  - Skilar síðustu tveimur línunum.
- tail +2 < inntak.txt
  - Sleppir fyrstu línunni.

Gerð N-stæðulíkana

Tvístæðulíkan

- Inntak: Málheild.
  1. Tilreiðing – eitt orð (tóki) í hverri línu.
  2. Búa til tvístæður: Skrifa út ord_i og ord_{i+1} í sömu línu.
  3. Talning.

Auðvelt í Unix/Linux

- G.r.f. að skrá *malheild.wrd* innihaldi einn tóka í línu.
- tail +2 < malheild.wrd > mailheild2.wrd
- paste mailheild.wrd mailheild2.wrd > malheild.bigrams
- sort malheild.bigrams | uniq -c | sort -nr > malheild.freq

Gerð N-stæðulíkana

Tvístæðulíkan

- Inntak: Málheild.
  1. Tilreiðing – eitt orð (tóki) í hverri línu.
  2. Búa til tvístæður: Skrifa út ord_i og ord_{i+1} í sömu línu.
  3. Talning.

Auðvelt í Unix/Linux

- G.r.f. að skrá *malheild.wrd* innihaldi einn tóka í línu.
- tail +2 < malheild.wrd > mailheild2.wrd
- paste mailheild.wrd mailheild2.wrd > malheild.bigrams
- sort malheild.bigrams | uniq -c | sort -nr > malheild.freq

Tvístæðulíkan í Perl (4.4.4 í kennslubók)

```
$text = <>;
while ($line = <>) { $text .= $line}
$text =~ tr /a-záðéíóýúæöþA-ZÁÐÉÍÓÝÚÆÖÞ0-9.,!?\-:;/\n/cs;
$text =~ s/([,.?!:;()\-])/\n$1\n/g;
$text =~ s/\n+/\n/g;
@words = split(/\n/, $text);

for ($i=0; $i<$#words; $i++) {
    $bigrams[$i] = $words[$i] . " " . $words[$i+1]; }

for ($i=0; $i <= $#bigrams; $i++) {
    if (!exists($frequency{$bigrams[$i]})) {$frequency{$bigrams[$i]} = 1;}
    else {$frequency{$bigrams[$i]}++;}
}
foreach $bigram (sort keys %frequency) {
    print "$frequency{$bigram} $bigram\n";
}
```

Outline

Tölfræðilíkön fyrir orðarunur

Sennileikalíkur (e. maximum likelihood)

- Lát $S = w_1, w_2, \ldots, w_n$ vera orðarunu.
- Með því að nota (þjálfunar)málheild (e. training corpus) M getum við áætlað líkurnar á þessari runu.
- $P(S)$ er hlutfallsleg tíðni strengsins S í M.
- $P(S)$ er kallað *maximum liklihood estimate* (MLE) fyrir S:

$$P_{MLE}(S) = \frac{C(w_1, w_2, \ldots, w_n)}{N} \quad (1)$$

N er heildarfjöldi strengja af lengd n í M.

Tölfræðilíkön fyrir orðarunur

Sennileikalíkur (e. maximum likelihood)

- Yfirleitt þó ómögulegt að finna þessa nálgun því stærð málheilda er endanleg!
- Einföldum því (1) og brjótum upp í:

$$P(S) = P(w_1, w_2, \ldots, w_n),$$

$$= P(w_1)P(w_2|w_1)P(w_3|w_1, w_2)\ldots P(w_n|w_1, \ldots, w_{n-1}),$$

$$= \prod_{i=1}^{n} P(w_i|w_1, \ldots, w_{i-1})$$

Tölfræðilíkön fyrir orðarunur

Takmarka þarf N-stæðurnar

- P(*It was a bright cold day in April*)
- P(S) = P(*It*) * P(*was*|*It*) * P(*a* | *It*, *was*) * P(*bright* | *It*, *was*, *a*) . . . * P(*April* | *It*, *was*, *a*, *bright* . . . , *in*)
- Í þessu dæmi þurfum við allt að áttstæður. Enginn málheild er svo stór að það gangi upp. Þessi líkindi eru því oftast nálguð (*Markov assumption*) með tvístæðum eða þrístæðum:

$$P(w_i|w_1, \ldots, w_{i-1}) \approx P(w_i|w_{i-1}) \quad (2)$$

$$P(w_i|w_1, \ldots, w_{i-1}) \approx P(w_i|w_{i-2}, w_{i-1}) \quad (3)$$

Tölfræðilíkön fyrir orðarunur

Líkur á setningu með tvístæðum og þrístæðum

$$P(S) = P(w_1) \prod_{i=2}^{n} P(w_i|w_{i-1})$$

$$P_{MLE}(w_i|w_{i-1}) = \frac{C(w_{i-1}, w_i)}{C(w_{i-1})}$$

$$P(S) = P(w_1)P(w_2|w_1) \prod_{i=3}^{n} P(w_i|w_{i-2}, w_{i-1})$$

$$P_{MLE}(w_i|w_{i-2}, w_{i-1}) = \frac{C(w_{i-2}, w_{i-1}, w_i)}{C(w_{i-2}, w_{i-1})}$$

Þjálfun, þróun og prófun

Mismunandi málheildir

- **Þjálfunarmálheild** (e. training corpus):
  - Málheild sem notuð er til ákvarða n-stæðurnar (mállíkanið).
- **Prófunarmálheild** (e. test corpus):
  - Málheild sem notuð er til að beita (prófa) mállíkaninu á.
- **Þróunarmálheild** (e. development corpus):
  - Málheild sem notuð er til að besta breytur sem líkanið notar.
- Allar þrjár málheildirnar þurfa að vera aðgreindar (frábrugðnar hver annarri).

n–fold cross-validation

- Málheild skipt upp í tvo hluta, þjálfunarmálheild og prófunarmálheild, með slembiaðferð.
- Mállíkanið lært af þjálfunarmálheild og beitt á prófunarmálheild.
- Endurtekið n-sinnum, ávallt með nýrri skiptingu fyrir þjálfunar- og prófunarmálheild.
- Meðaltal tekið af niðurstöðum.
- Kallað 10–*fold cross-validation* þegar $n = 10$.
- Gerir það að verkum að niðurstöður eru ekki háðar einni upphaflegri skiptingu í þjálfunar- og prófunarmálheild.

Orðasafn, orðaforði (e. vocabulary)

- Orð sem ekki eru hluti af mállíkaninu (þ.e. hafa ekki fundist í þjálfunarmálheildinni) munu koma fyrir í prófunarmálheild.
- Af hverju er það næstum öruggt?
- Þessi orð eru kölluð “óþekkt” (e. unknown eða e. out-of-vocabulary (OOV)).
- Jafnframt er tíðni sjaldgæfra orða ekki áreiðanleg.
- Tvær aðferðir til að meðhöndla óþekkt orð:
  - *Closed vocabulary*. Óþekkt orð hunsuð.
  - *Open vocabulary*. Sérstakar ráðstafanir gerðar til að meðhöndla óþekkt orð, t.d. sléttun.

Naum gögn (e. sparse data)

- Mállíkön byggja á málheildum sem eru ekki nógu stórar til að ákvarða tíðnitölur fyrir allar tví- og þrístæður með áreiðanlegum hætti.
- Orðasafn: 20.000 orðmyndir.
  - Tvístæður: $20.000^2 = 400.000.000$
  - Þrístæður: $20.000^3 = 8.000.000.000.000$
- Þjálfunargögnin eru því naum. Margar n-stæður fá núll líkindi sem er óraunsætt (sjá dæmi bls. 99).
- MLE aðferðin tekur ekki á núll líkindum.
- Þurfum því aðferð til að ákvarða líkindi á "óséðum" n-stæðum => Sléttun (e. smoothing)

Outline

Sléttun

Regla Laplace (1820)

- Einfaldlega bætir einum við allar tíðnitölur.
- Þess vegna oft kallað “the add one method”.
- Tíðni óséðra n-stæða er þá 1.

$$P_{Laplace}(w_{i+1}|w_i) = \frac{C(w_i, w_{i+1}) + 1}{C(w_i) + Card(V)}$$

$Card(V)$ er stærð orðasafnsins (fjöldi orðmynda).

Sléttun – tafla bls. 100

w_i, w_{i+1}	$C(w_i, w_{i+1})$	$C(w_i) + Card(V)$	$P_{Lap}(w_{i+1}\|w_i)$
<s>a	133	7072 + 8634	0.008500
a good	14	2482 + 8634	0.001300
good deal	0	53 + 8634	0.000120
deal of	1	5 + 8634	0.000230
of the	742	3310 + 8634	0.062000
the literature	1	6248 + 8634	0.000130
literature of	3	7 + 8634	0.000460
of the	742	3310 + 8634	0.062000
the past	70	6248 + 8634	0.004800
past was	4	99 + 8634	0.000570
was indeed	0	2211 + 8634	0.000092
indeed already	0	17 + 8634	0.000120
already being	0	64 + 8634	0.000110
...			
this way	3	264 + 8634	0.000450

Table: Tíðni tvístæða með Laplace sléttun.

Sléttun

Gallar við reglu Laplace

- Óséðar n-stæður fá mikinn "tölfræðilegan massa".
  - Hin ólíklega tvístæða *the of* fær t.d. tíðnina 1, fjórðung af tíðni (algengu) tvístæðunnar *this way*.
- *Discount factor* er hlutfallið á milli MLE líkinda og líkindanna eftir sléttun. Þetta gildi vill oft verða of hátt þegar Laplace er notað.
- Dæmi:
  - Samkvæmt mállíkaninu er MLE líkindi fyrir *this way* = $\frac{3}{264} = 0.0114$. Eftir sléttun eru líkindin 0.00045.
  - *Discount value* er þá: $\frac{0.0114}{0.00045} = 24.4$.
  - MLE líkindin fyrir þessa tvístæðu hefur sem sagt verið "discounted by" 24.4.

Sléttun

Good-Turing estimation (1953)

- Ein skilvirkasta aðferðin við sléttun.
- Endurmetur tíðni n-stæða, sem fundist hafa í málheild, með því að lækka hana og færa tölfræðilegan massa yfir á óséðar n-stæður (á sama hátt og Laplace regla gerir).
- *Discount factor* er hins vegar breytilegur og háður því hversu oft tiltekin n-stæða finnst í málheildinni.

Skilgreining

- Lát N_c vera fjölda n-stæða sem koma fyrir nákvæmlega c sinnum í málheildinni.
- N_0 er fjöldi af óséðum n-stæðum, N_1 er fjöldi af n-stæðum sem koma fyrir einu sinni, o.s.frv.

Sléttun

Good-Turing estimation (1953)

- Ein skilvirkasta aðferðin við sléttun.
- Endurmetur tíðni n-stæða, sem fundist hafa í málheild, með því að lækka hana og færa tölfræðilegan massa yfir á óséðar n-stæður (á sama hátt og Laplace regla gerir).
- *Discount factor* er hins vegar breytilegur og háður því hversu oft tiltekin n-stæða finnst í málheildinni.

Skilgreining

- Lát N_c vera fjölda n-stæða sem koma fyrir nákvæmlega c sinnum í málheildinni.
- N_0 er fjöldi af óséðum n-stæðum, N_1 er fjöldi af n-stæðum sem koma fyrir einu sinni, o.s.frv.

Good-Turing

- Endurákvarðar tíðni n-stæða sem koma fyrir c sinnum með formúlunni:

$$c* = (c+1)\frac{N_{c+1}}{N_c}$$

- Fyrir óséðar n-stæður: $c* = \frac{N_1}{N_0}$
- Fyrir n-stæður sem komu einu sinni fyrir: $c* = \frac{2*N_2}{N_1}$
- Fyrir n-stæður sem komu tvisvar sinnum fyrir: $c* = \frac{3*N_3}{N_2}$
- Skilyrðalíkurnar eru þá:

$$P_{GT}(w_n|w_1,\ldots,w_{n-1}) = \frac{c^*(w_1,\ldots,w_n)}{C(w_1,\ldots,w_{n-1})}$$

Good-Turing – tafla bls. 102

Frequency of occurrence	N_c	c*
0	74.523.701	0.0005
1	37.365	0.31
2	5.820	1.09
3	2.111	2.02
4	1.067	3.37
5	719	3.91
6	468	4.94
...		

Table: Tíðni tvístæða með Good-Turing sléttun.

Takið eftir að hin endurákvarðaða tíðni tvístæða sem sáust ekki í málheildinni er aðeins 0.0005.

Línuleg brúun

- *Linear interpolation*, *Deleted interpolation*
- Sameinar línulega *MLE* af lengd 1 til n.
- Nálgunin fyrir sérhverja óséða n-stæðu er háð þeim orðum sem hún inniheldur.
- Fyrir þrístæður:
  $P_{DelInterpolation}(w_n|w_{n-2}, w_{n-1}) =$
  $\lambda_1 P_{MLE}(w_n|w_{n-2}, w_{n-1}) +$
  $\lambda_2 P_{MLE}(w_n|w_{n-1}) + \lambda_3 P_{MLE}(w_n)$
- þar sem $0 \leq \lambda_i \leq 1$ og $\sum_{i=1}^{3} \lambda_i = 1$
- Hægt að þjálfa og besta λ_i með því að nota málheild.