## Statistical Identification of Language

$$\Gamma(n) = \int_0^\infty t^{n-1} e^{-t}\,dt$$

$$\Gamma'(n) = \frac{d\int_0^\infty t^{n-1}e^{-t}\,dt}{dn} = \int_0^\infty t^{n-1}e^{-t}\log t\,dt$$

$$\mathcal{L}(t^n) = \frac{\Gamma(n+1)}{s^{n+1}}$$

$$\mathcal{L}(\log t) = \frac{(\gamma - \log s)}{s}$$

- Inngangur
- Fyrri rannsóknir
  - Einstakar stafarunur
  - Algeng orð
  - N-stæðu módel og tölfræðileg röðun
- "Our" Classification Methods
  - Markov módel
  - Bayesian Decision Rules
  - Val á parametrum módelsins
- Niðurstöður
  - Lítil þjálfunarmálheild (mynd 1)
  - Stór þjálfunarmálheild (mynd 2)
  - Textagögn og Genagögn
- Lokaspjall

## Inngangur

- Horfum á nokkra strengi sem eru 20 stafir á lengd.
  Þeir eru valdir blindandi af netinu með því að veifa músinni um skjáinn.
  Sjáið þið hvaða tungumál þetta eru ?
  - `blundede bare et hal`
  - `evure chimique (5 g)`
  - `er Protestler träumt`
- Hversu flókið er að skrifa forrit sem þekkir þá?
- Þarf ég að kunna frönsku?
- Hvað þarf langan streng til að málið þekkist?
- Og til hvers í ósköpunum ættum við að gera þetta?

## Inngangur (frh.)

- **Dunning & co**. skrifuðu x00 lína C forrit þar sem voru 50 tölfræðimódel
- Byggir ekki á málfræðiþekkingu eða handskrifuðum reglum
- Forritið lærir af því að lesa texta "Data driven" <- þessi lína er 50 stafir
- Virkar~ á 10 stafa strengi og mjög vel á 50+ stafa strengi
- Þjálfunartextinn þarf ekki að vera nema nokkur þúsund orð
- Gagnlegt við:
  - Sjálfvirkar þýðingar: til að fatta ef hluti af texta er latína/franska ..
  - Google þykist vita á hvaða máli síður eru
- Aukaafurð:
  - Ef forritið er matað á þýskum texta, án þess að hafa séð hann áður þá flokkar það hann sem enskan. -> Finnur skyldleika mála, búa til ættartré
  - Sömu hugmynd má taka lengra og nota forritið til að þekkja genaraðir (!) til að staðfesta að frumusýni sé úr manni eða til að finna skyldleika tegunda

## Fyrri rannsóknir

- Spjall
  - Ekki til kerfisbundinn samanb á aðferðum sem hafa verið reyndar
  - Margar aðferðir byggja á þekkingu á málum (linguistic knowledge)
  - Sumar gera ráð fyrir að málinu sé skipt í orð (hvað með kínversku?)
  - Verkefnið er einfalt m.v. sjálfvirkar þýðingar og gagnaöflun af netinu
- Rannsóknir
  - [Chu94] Unique letter combinations (e. –ery, f. –eux, i. –cchi, d. der)
  - [Joh93] Common words, virkar vel ef strengurinn er nógu langur
  - N-gram counting using rank order statistics (short char seq's)
    - Tilreiða og telja stuttar n-stæður (stafa, ekki orða)
    - Ekki ólíkt rannsókninni í þessari grein

## Ted Dunnings' aðferð:

- Les þjálfunarmálheildir staf fyrir staf til að þróa **Markov** tölfræði módel fyrir þau mál sem þjálfað er fyrir.
- í grunninn byggja módelin á talningu á n-stæðum

| 1-stæður | 2-stæður | 3-stæður |
|---|---|---|
| E 245 | EA 13 | EAN 6 |
| A 167 | AN 16 | EAR 3 |
| O 156 | AR 15 | AEN 2 |
| T 153 | TA 11 | ETA 1 |

- Notar **Markov-módel**ið til að meta líkurnar á hverjum streng fyrir sig og svo **Bayes' Decision rules** til að ákvarða líklegasta málið

## Markov Models

Markov grfa líkurnar á næstu stöðu byggi bara á núverandi stöðu.

- $p(S) = p(s_1 s_2 s_3 .. s_n) = p(s_1) \prod_{i=2}^{} p(s_i | s_{i-1})$ (1-stæða, 2-stæða)

Dæmi:

- p(EAN)= p(E)*p(A|E)*p(N|A)
- s er 1 stafur -> order 1 módel. Notar 1- og 2-stæður
- Einnig hægt að líta á stöðu sem lengri streng
- p(EAN)= p(E)*p(A|E)*p(N|EA)
- Her er s 2 stafir -> order 2 módel. Notar einnig 3-stæður

## Bayesian Decision Rules

**2 kökudiskar**

- D#1 (10x o 30x ○) ○○○o○○○o○○○o○○○o○○○o○○○o○○○o○○○o○○○o○○○o○○○o
- D#2 (20x o 20x ○) ○o○o○o○o○o○o○o○o○o○o○o○o○o○o○o○o○o○o○o○o○o○o
- Vel eina köku af handahófi og hún er ○
- Hverjar eru líkurnar á að hún hafi komið af D#1? > 0.5

- Skv. Bayes
- $P(D\#1|○) = \frac{P(○|D\#1) * P(D\#1)}{P(○)} = \frac{0.75 * 0.5}{50/80} = 0.6$

- Á sama hátt segjum við.. nú fékk ég þennan streng ○, hvaða líkur eru þá á því að hann sé úr spænsku D#1 ? (mál-módelin gefa okkur P(○|D#1))

## Niðurstöður

**Litið þjálfunarsett 2K (bls í Word)**

Order of Model
x-ásinn merkir hvaða n-stæður voru notaðar í Markov módelinu.

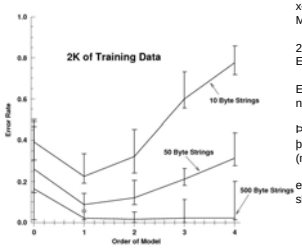
2-3-stæður reynast best (order 1-2)
Error rate ~ 0.0 - 0.3 (miðgildi mean)

Efsta línan sýnir 10 stafa strengi
nær að flokka um 70% af þeim rétt !

Þó error rate sé lágt fyrir 500 stafi,
þá breikka vikmörkin með hærri gráðu
(módelið á engin dæmi um 5-stæðurnar .. )

ef forritið vissi ekkert, þá var ekki giskað heldur skilað villu

Skoða mynd ..




## Niðurstöður

**Stór þjálfunarmálheild 50K (jafnast 20-30 bls í Word)**

Módel af hærri gráðu verða góð
enda eru þá líkur á að 4-stæðurnar og 5-stæðurnar hafi sést í þjálfun

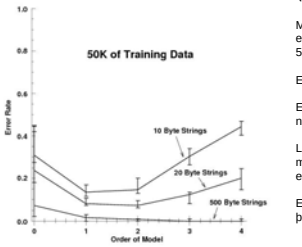
Error rate ~ 0.0 - 0.15 (miðgildi, mean)

Efsta línan sýnir 10 stafa strengi
nær að flokka 85% af þeim rétt !

Langir textastrengir þekkjast fullkomlega með módelum af gráðu 2 og hærra. Og vikmörkin eru nánast engin.

En vikmörkin fyrir 0-gráðu módelið eru há, þó verið sé að vinna með langa strengi

Skoða mynd ..



## Gögnin

- Í ljós kom að **máli skiptir**:
  - Hvernig prófunar strengir eru valdir
  - Stærð þjálfunarmálheildar
  - Stærð strengjanna sem á að þekkja
  - Fjöldi málanna sem þarf að greina á milli
  - Hvort það er fylgni milli máls og umfjöllunarefnis
    forritið greinir sundur efnisflokka rétt eins og tungumál

## Gögnin

- Texta-málheildir
  - Notuðu sama texta þýddan milli ensku og spænsku (parallel corpus)
    til að vera með samskonar efnisval í báðum málum.
  - 50 þjálfunartextar valdar random, og 600 prófunarstrengir úr hvoru máli.
  - Þeir hentu prófunarstrengjum sem innihéldu t.d. aðallega bil, formúlur eða tákn
    eða ef strengurinn úr spænska hlutanum var á ensku og öfugt.
  - Hægt er að nálgast prófunarmálheildin þeirra frá `lexical@nmsu.edu`
- Genamengið
  - Basa-raðir `[GATTTCGAAAT...]` úr þremur mismunandi lífverum
  - Ekki hægt að skoða prófunarstrengina í höndunum `:]`
  - Bara 1 þjálfunartexti fyrir hverja lífveru.
  - Prófunarstrengirnir eru mikið lengri
  - Parallel corpus hefur ekki augljósa merkingu milli manns og gerils

## Lokaspjall

- Ef þjálfunartextinn er stuttur, þá virka low-order Markov módel best
- `- - - - - - -` langur, `- - -` high `- - - - - - - - -`
- 50K þjálfun nær 92% árangri með 20 stafi, en 99.9% með 500 stafi
- Rannsóknin á lífsýnabankanum leiddi í ljós að mörg sýnin sem áttu að vera úr mönnum reyndust vera menguð og þarmeð ónothæf.
- Gæti verið betri nálgun á að þekkja talað mál, en aðferðir sem byggja á tokenization, Ekki viðkvæmt fyrir noise
- Væri forvitnilegt að nota þetta til að smíða ættartré mála sem byggja á þeim málýskum sem notaðar eru en ekki eingöngu "há-þýsku" og norðlensku.
- Nothæft til að koma í veg fyrir að óhrein sýni fari inní lífsýnabanka

**Takk fyrir**

Any ?s