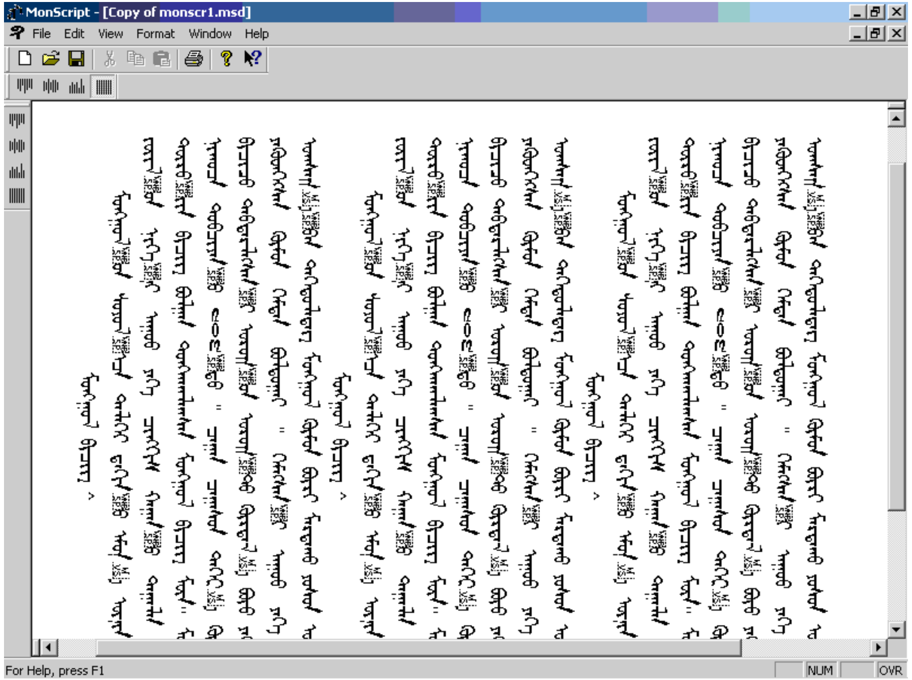


Хэлний загварчлал

Дэвшилтэт Good Turing тэгшлэлт

Санамж: Нэгийг нэмэх (Лапласын) тэгшлэлт

$$P_{Add\text{-}1}(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i) + 1}{c(w_{i-1}) + V}$$

Илүү ерөнхий томьёолол: k нэмэх

$$P_{Add-k}(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i) + k}{c(w_{i-1}) + kV}$$

$$P_{Add-k}(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i) + m(\frac{1}{V})}{c(w_{i-1}) + m}$$

Юниграм өмнөх рүү тэгшлэлт

$$P_{Add-k}(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i) + m(\frac{1}{V})}{c(w_{i-1}) + m}$$

$$P_{\mathrm{UnigramPrior}}(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i) + mP(w_i)}{c(w_{i-1}) + m}$$

Дэвшилтэт тэгшлэх алгоритмууд

- Олон тэгшлэлтийн алгоритм ашиглагдаж байна
  - Good-Turing
  - Кнессер-Ней (Kneser-Ney)
  - Виттэн-Бэл (Witten-Bell)
- **Нэг удаа таарсан** үгийн тоог ашиглан
  - **Хэзээ ч таараагүй** үгийн тоог үнэлэхэд тусалдаг

Тэмдэглэгээ: N_c = с дамтамжийн дамтамж

- N_c = с удаа таарсан үгсийн тоо
- Sam I am I am Sam I do not eat

```
I   3
sam 2
am  2
do  1
not 1
eat 1
```

$N_1 = 3$ (do, not, eat)

$N_2 = 2$ (sam, am)

$N_3 = 1$ (I)

Good-Turing тэгшлэлтийн төсөөлөл

- Загасчлал (Жош Гудмены жишээ):
  - 10 мөрөг, 3 алгана, 2 цагаан, 1 хулд, 1 яргай, 1 могой загас = 18 загас
- Дараагийнх нь яргай загас байх магадлал хэр вэ?
  - 1/18
- Дараагийнх нь өөр загас байх магадлал хэр вэ? (ж.нь. Муур загас)
  - Шинэ зүйлийг үнэлэх нэг удаа харсан зүйлийн үнэлгээг ашиглацгаая.
  - 3/18 (учир нь N_1=3)
- Ийм гэж үзвэл, дараагийнх нь яргай байх магадлал нь ... гэж
  - 1/18 –аас бага байх ёстой
  - Үүнийг хэрхэн үнэлэх вэ?

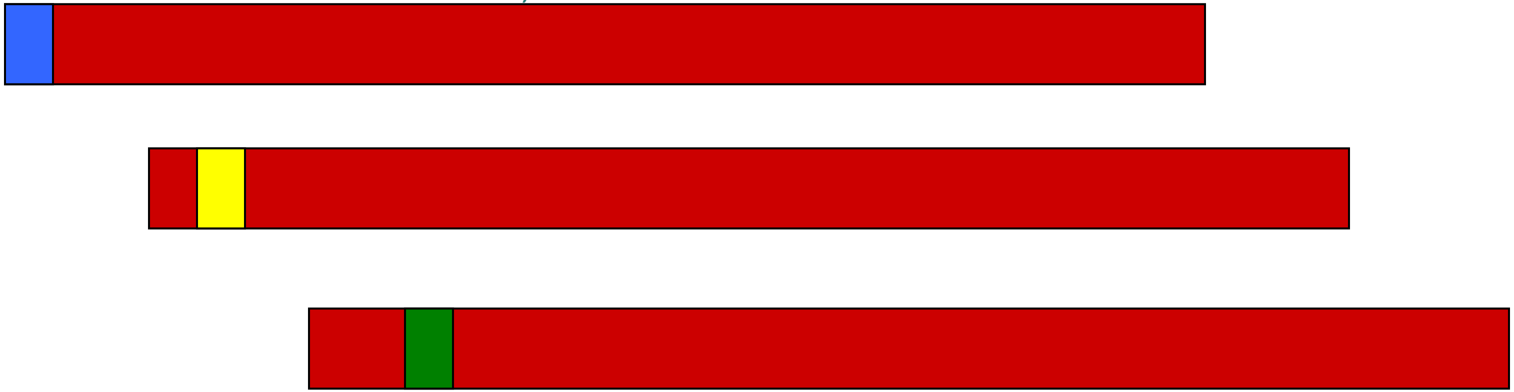
Good Turing бодолтууд

$$P^*_{GT}(\text{things with zero frequency}) = \frac{N_1}{N} \qquad c^* = \frac{(c+1)N_{c+1}}{N_c}$$

- Тааraагүй (муур загас)
  - c = 0:
  - MLE p = 0/18 = 0
  - P^*_{GT} (таараагүй) = N_1/N = 3/18
- Нэг удаа таарсан(яргай)
  - c = 1
  - MLE p = 1/18
  - C*(яргай) = 2 * N2/N1
    = 2 * 1/3
    = 2/3
  - P^*_{GT}(яргай) = 2/3 / 18 = 1/27

Нейнийн багийнхны Good Turing-ийн төсөөлөл

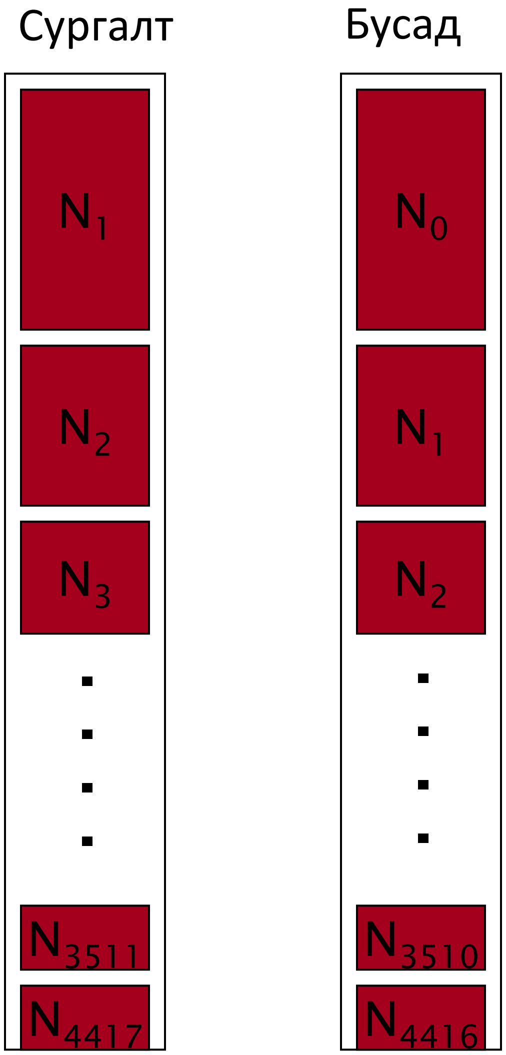
H. Ney, U. Essen, and R. Kneser, 1995. On the estimation of 'small' probabilities by leaving-one-out. IEEE Trans. PAMI. 17:12,1202-1212



Бусад үгс:

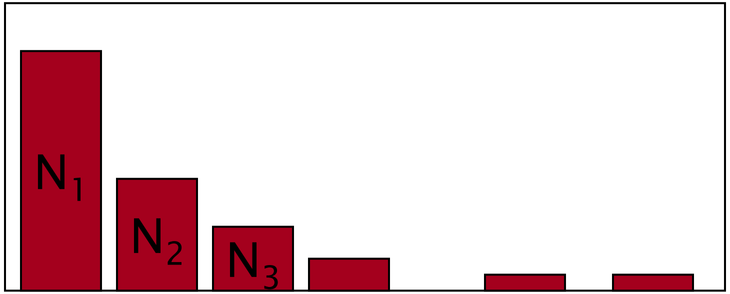
Нейнийн багийнхны Good Turing-ийн төсөөлөл

- Нэгийг хасах үнэлгээний санаа
  - *c* тооны сургалтын үгсийг тус бүрд нь ээлжлэн авч үзье
  - *c* сургалтын олонлогийн хэмжээ *c*–1, бусад үгсийн олонлогийн хэмжээ 1
  - Сургалтанд харагдаагүй бусад үгсийн хувь хэмжээ ямар байх вэ?
    - N_1/c
  - Сургалтанд *k* удаа харагдсан бусад үгсийн хувь хэмжээ ямар байх вэ?
    - $(k+1)N_{k+1}/c$
  - Иймээс ирээдүйд сургалтын тоо *k* байх үгсийг $(k+1)N_{k+1}/c$ магадлалтай хүлээнэ
  - Энд сургалтын тоо k бүхий N_k үгс байна
  - Тус бүр дараах магадлалтай харагдах ёстой:
    - $(k+1)N_{k+1}/c/N_k$
  - …эсвэл хүлээгдэж буй тоотой: $k^* = \frac{(k+1)N_{k+1}}{N_k}$

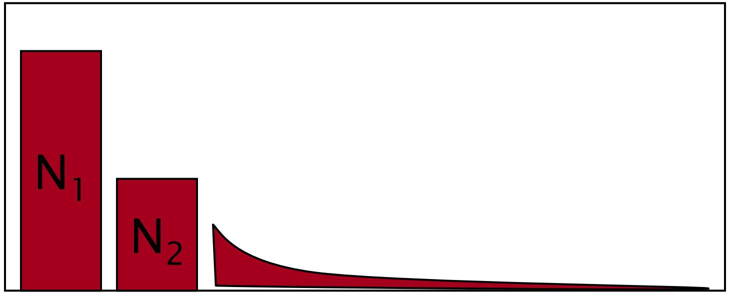


Good-Turing –ийн хүндрэл

- Асуудал: “the” яах вэ? (хэлэгдсэн c=4417)
  - бага k –ийн хувьд, $N_k > N_{k+1}$
  - их k –ийн хувьд, хэт зөрөөтэй, тэгүүд нь үнэлгээг будлиулаад байна



  - Энгийн Good-Turing [Гейл болон Сэмпсон]: Туршилтын N_k –н найдваргүй утгуудыг хувирлын хамгийн сайн тохирох хууль дүрмээр солих

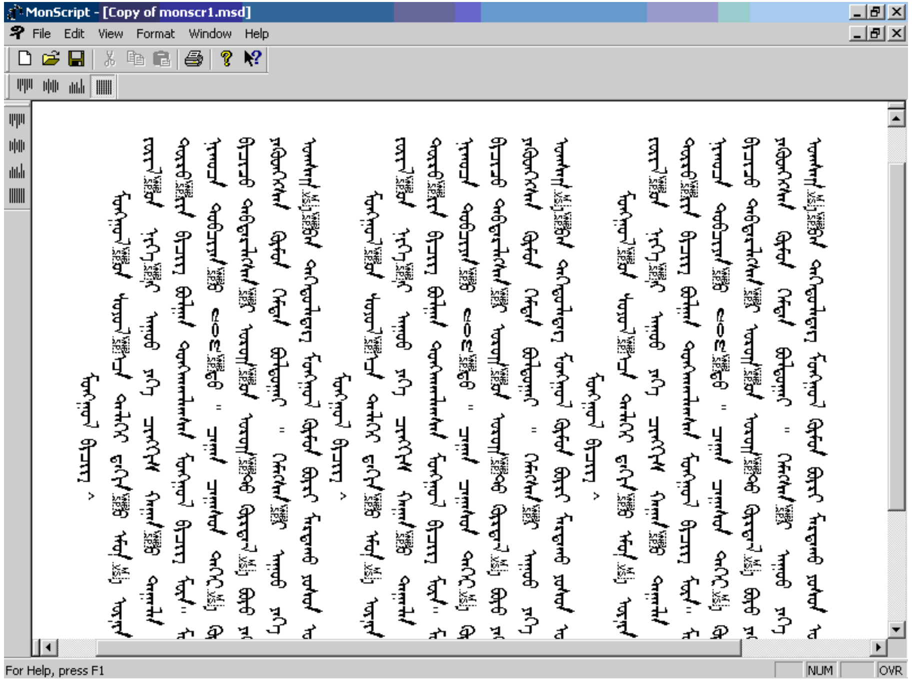


Good-Turing –ийн тоонуудын дүгнэлт

- Чиэрч болон Гейл(1991) –ийн тоонууд
- Associated Pres News –ийн 22 сая үгс

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

Count c	Good Turing c*
0	.0000270
1	0.446
2	1.26
3	2.24
4	3.24
5	4.22
6	5.19
7	6.21
8	7.24
9	8.25



Хэлний загварчлал

Дэвшилтэт Кнессер-Ней тэгшлэлт

Good-Turing тоог дүгнэвэл

- Чиэрч болон Гейл (1991)-н тоо
- АП –ийн мэдээний сангийн 22 сая үг

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

- Энэ нь дараах байдалтай харагдаж байна. c* = (c - .75)

тоо c	Good Turing c*
0	.0000270
1	0.446
2	1.26
3	2.24
4	3.24
5	4.22
6	5.19
7	6.21
8	7.24
9	8.25

Абсолют бууруулах интерполяци

- Тооцооллыг хэмнэж, зөвхөн 0.75-г хасах (эсвэл d-г хасах)!

Бууруулах биграм

Интерполяцийн жин

$$P_{\mathrm{AbsoluteDiscounting}}(w_i \mid w_{i-1}) = \frac{c(w_{i-1}, w_i) - d}{c(w_{i-1})} + \lambda(w_{i-1})P(w)$$

Юниграм

- (тоо 1 ба 2-ын хувьд нэмэлт хос утгуудыг хадгалж магадгүй)
- Гэвч зөвхөн энгийн P(w) юниграмыг ашигласан нь дээр үү?

Кнессер-Ней тэгшлэлт I

- Бага эрэмбэтэй юниграмын магадлалын хувьд сайн үнэлэх!
  - Шаноны тоглоом: *I can't see without my reading* Francisco glasses ___________?
  - Юниграмд "Francisco" нь "glasses" –ийг бодвол илүү нийтлэг
  - … гэвч "Francisco" ихэвчлэн "San" –гийн дараа байдаг. Мөн "san-ээс олон давтагддаг.
- Энэ нь биграмд харагдаагүй бол юниграм чухал хэрэгтэй!
- P(w)-ийн оронд: "w хэр магадлалтай вэ"
- $P_{continuation}(w)$: "дараа үргэлжлэх үг (novel continuation) байдлаар w харагдах магадлал хэд вэ?
  - Үг бүрийн хувьд, түүнийг гүйцээдэг биграмын төрлийг тоолох
  - Биграмын төрөл бүр анх харагдаж эхэлсэн үедээ novel continuation байдаг.

$$P_{CONTINUATION}(w) \propto \left|\{w_{i-1} : c(w_{i-1}, w) > 0\}\right|$$

Кнессер-Ней тэгшлэлт II

- Хэдэн удаа w нь novel continuation байдлаар харагдсан вэ:

$$P_{CONTINUATION}(w) \propto \left|\{w_{i-1} : c(w_{i-1}, w) > 0\}\right|$$

- Нийт биграм төрлүүдийн тоогоор нормчловол

$$\left|\{(w_{j-1}, w_j) : c(w_{j-1}, w_j) > 0\}\right|$$

$$P_{CONTINUATION}(w) = \frac{\left|\{w_{i-1} : c(w_{i-1}, w) > 0\}\right|}{\left|\{(w_{j-1}, w_j) : c(w_{j-1}, w_j) > 0\}\right|}$$

Кнессер-Ней тэгшлэлт III

- 2 дахь метафор: w – ийн өмнө харагддаг үгийн бүх төрлийн нийлбэр тоо

$$|\{w_{i-1} : c(w_{i-1}, w) > 0\}|$$

- Тухайн үгийн өмнөх орших болох бүх үгсийн биграм хосын тооны нийлбэрээр нормчлох:

$$P_{CONTINUATION}(w) = \frac{|\{w_{i-1} : c(w_{i-1}, w) > 0\}|}{\sum_{w'} |\{w'_{i-1} : c(w'_{i-1}, w') > 0\}|}$$

- (Francisco) нь зөвхөн нэг удаа (San)-гийн дараа тохиолдвол энэ нь үргэлжлэх магадлал бага гэсэн үг.

Кнессер-Ней тэгшлэлт IV

$$P_{KN}(w_i \mid w_{i-1}) = \frac{\max(c(w_{i-1}, w_i) - d, 0)}{c(w_{i-1})} + \lambda(w_{i-1}) P_{CONTINUATION}(w_i)$$

λ бол нормчлох тогтмол; хассан магадлалын хэмжээ

$$\lambda(w_{i-1}) = \frac{d}{c(w_{i-1})} \left| \{ w : c(w_{i-1}, w) > 0 \} \right|$$

Нормчилсон хасалт

w_{i-1} –г дагаж болох үгийн төрлүүдийн тоо
= хассан үгийн төрлүүдийн тоо
= нормчилсон хасалтыг хэрэгжүүлсэн удаагийн тоо

Кнессер-Ней тэгшлэлт: Рекурс томьёолол

$$P_{KN}(w_i \mid w_{i-n+1}^{i-1}) = \frac{\max(c_{KN}(w_{i-n+1}^{i}) - d, 0)}{c_{KN}(w_{i-n+1}^{i-1})} + / (w_{i-n+1}^{i-1}) P_{KN}(w_i \mid w_{i-n+2}^{i-1})$$

$$c_{KN}(\bullet) = \begin{cases} count(\bullet) & \text{Хамгийн өндөр эрэмбийн хувьд} \\ continuationcount(\bullet) & \text{Бага эрэмбийн хувьд} \end{cases}$$

Үргэлжлэх тоо = •-н хувьд давтагдашгүй ганц үгийн орчны тоо