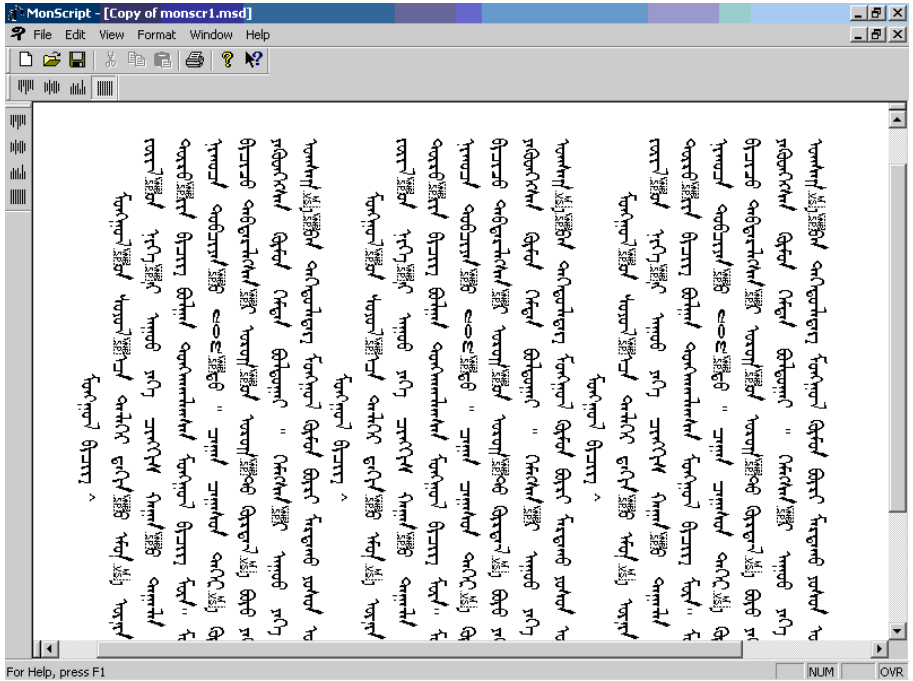


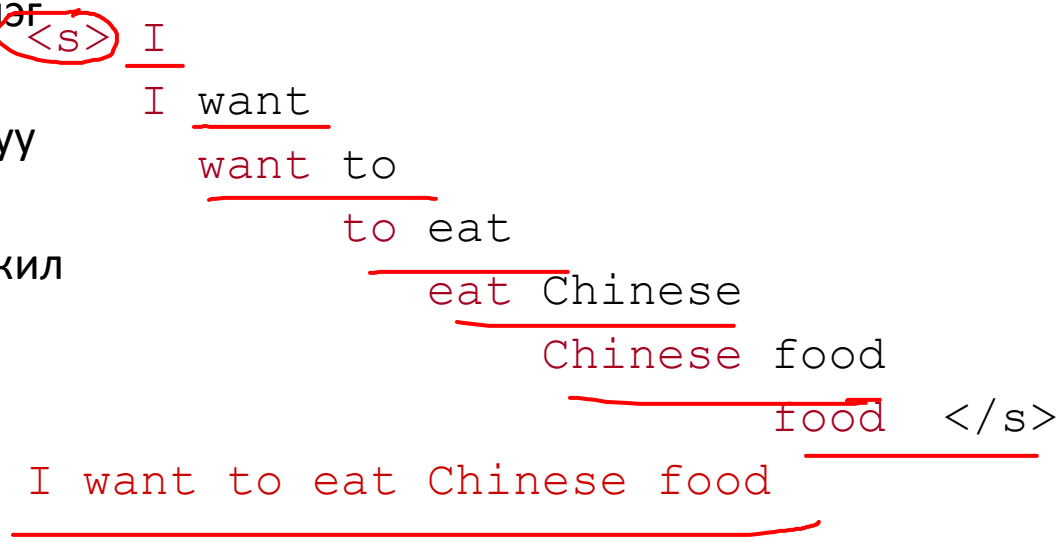
Хэлний загварчлал

Ерөнхийлөл ба тэгүүд

2021 он

Шаноны дүрслэх арга

- түүний (<s>, w) магадлалын дагуу нэг санамсаргүй биграм сонго
- одоо түүний (w, x) магадлалын дагуу нэг санамсаргүй биграм сонго
- Ийм замаар </s> -г сонготол үргэлжил
- Дараа нь үсгийг хооронд нь холбо



Шекспирыг ерөнхийлбэл

Unigram

To him swallowed confess hear both. Which. Of save on trail for are ay device and rote life have
Every enter now severally so, let
Hill he late speaks; or! a more to leg less first you enter
Are where exeunt and sighs have rise excellency took of.. Sleep knave we. near; vile like

Bigram

What means, sir. I confess she? then all sorts, he is trim, captain.
Why dost stand forth thy canopy, forsooth; he is this palpable hit the King Henry. Live king. Follow.
What we, hath got so she that I rest and sent to scold and nature bankrupt, nor the first gentleman?

Trigram

Sweet prince, Falstaff shall die. Harry of Monmouth's grave.
This shall forbid it should be branded, if renown made it empty.
Indeed the duke; and had a very good friend.
Fly, and will rid me these news of price. Therefore the sadness of parting, as they say, 'tis done.

Quadrigram

King Henry.What! I will go seek the traitor Gloucester. Exeunt some of the watch. A great banquet serv'd in;
Will you not tell me who I am?
It cannot be but so.
Indeed the short and the long. Marry, 'tis a noble Lepidus.

Шекспирын корпус

- N=884,647 токен, V=29,066 үгийн сан
- V^2= 844 сая боломжит биграмаас 300,000 биграм төрлүүд олдсон.
  - Иймээс боломжит биграмуудын 99.96% нь тааралдаагүй (хүснэгтэнд тэг бичилт)
- Квадриграм үүсэх боломж муу: Шекспирийн зохиол нь өөрөө гарч ирнэ. Учир корпус нь бага.

Валл стритын сэтгүүл бол Шекспир биш

Unigram

Months the my and issue of year foreign new exchange's september were recession exchange new endorsed a acquire to six executives

Bigram

Last December through the way to preserve the Hudson corporation N. B. E. C. Taylor would seem to complete the major central planners one point five percent of U. S. E. has already old M. X. corporation of living on information such as more frequently fishing to keep her

Trigram

They also point to ninety nine point six billion dollars from two hundred four oh six three percent of the rates of interest stores as Mexico and Brazil on market conditions

Хэт тааруулах аюул

- Хэрэв тексийн корпус сургалтын өгөгдөл шиг харагддаг бол N-грам зөвхөн үг таахад сайн ажилладаг
  - Бодит амьдралд, үргэлж ийм байдаггүй
  - Ерөнхий тохиолдолд сайн ажиллах зөв загварыг сургах хэрэгтэй!
  - Ганцхан төрлийн ерөнхийлөл: Тэгд хүргэнэ!
    - Сургалтын олонлогт байхгүй
      - Боловч тестийн олонлогт таарна

Тэгүүд

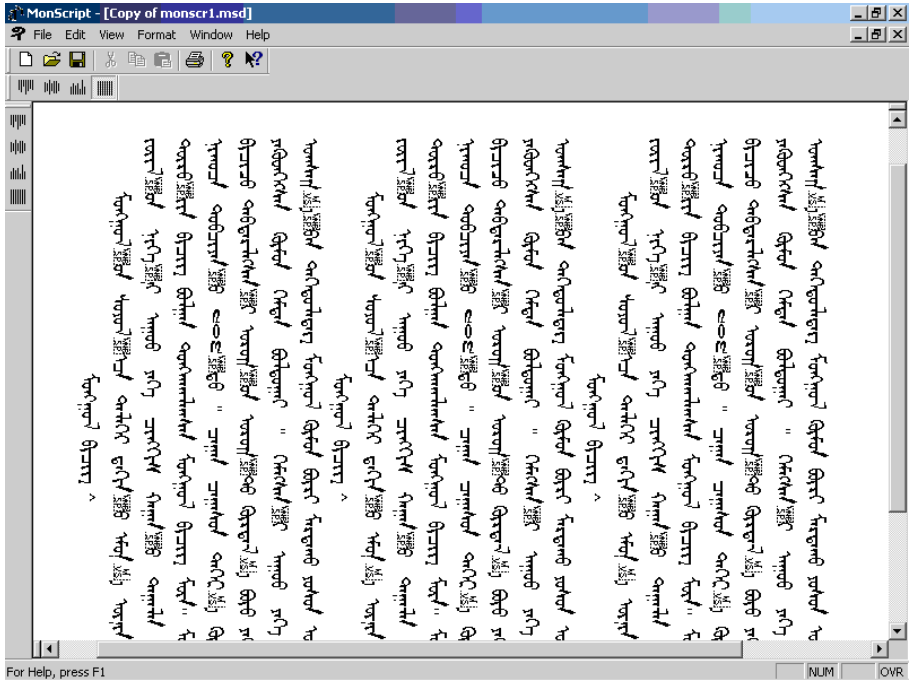
- Сургалтын олонлог:
  … denied the allegations
  … denied the reports
  … denied the claims
  … denied the request

- Тестийн олонлог
  … denied the offer
  … denied the loan

P(“offer” | denied the) = 0

Тэг магадлалтай биграм

- Тэг магадлалтай биграм
  - Тестийн олонлогт 0 магадлал оноох!
- Иймээс эргэлзээт чадварыг тооцоолж чадахгүй (тоог тэгд хувааж болохгүй)



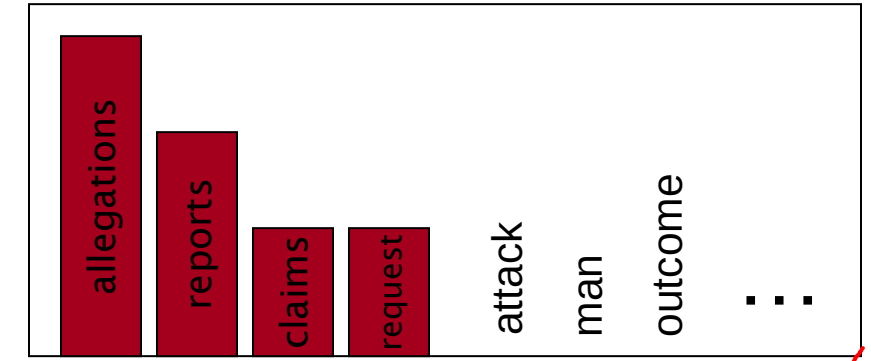
Хэлний загварчлал

Тэгшлэлт: Нэгийг нэмэх (Лапласын) тэгшлэлт

Тэгшлэх төсөөлөл (Дэн Клейн зохиосон)

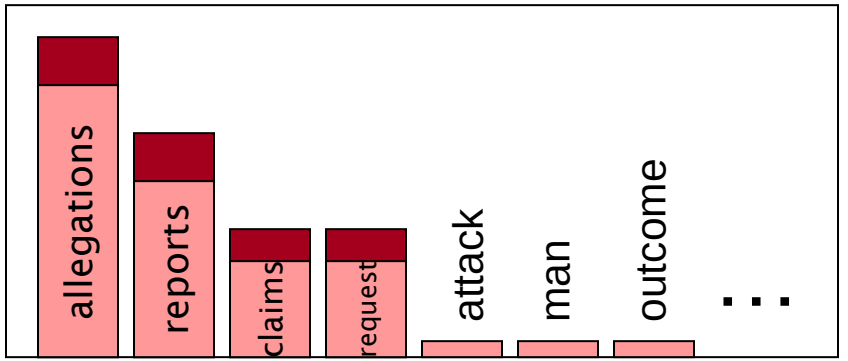
- Сарнисан статистик байхад:

P(w | denied the)
3 allegations
2 reports
1 claims
1 request
7 total



- Бусад руу тараавал тэгүүд алга болно

P(w | denied the)
2.5 allegations
1.5 reports
0.5 claims
0.5 request
2 other
7 total



Нэгийг нэмэх үнэлгээ

smoothing

- Мөн Лапласын тэгшлэлт гэдэг
- Үг бүр өмнөхөөсөө нэгээр илүү болж байгаа мэт харагдана
- Бүх тоон дээр зөвхөн нэг нэмнэ!

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

- ХамИхҮнэХувийн үнэлгээ:

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

үгсийн сан

- Нэгийг нэмэх үнэлгээ:

Хамгийн их үнэний хувийг үнэлэх

- Хамгийн их үнэний хувийг үнэлэх
  - Т сургалтын олонлогоос М загварын зарим параметрүүд
  - Өгөгдсөн М загварын Т сургалтын олонлог дээр үнэний хувийг хамгийн их болгох
- Нэг сая үгтэй корпуст “торх” гэдэг үг 400 удаа тохиолддог гэж үзье
- Дурын үгийг авахад “торх” гэдэг үг байх магадлал хэд вэ?
- ХамИхҮнэХувийн үнэлгээгээ 400/1,000,000 = .0004
- 400-аас олон тохиолддог корпусын хувьд энэ бага үнэлгээ
  - Гэхдээ сая үгэнд 400 тохиолдоно гэдэг **их магадлалтай** гэж **үнэлж** байна.

Рестораны корпус: Лапласын тэгшилсэн биграм тоо

	i	want	to	eat	chinese	food	lunch	spend
i	6	828	1	10	1	1	1	3
want	3	1	609	2	7	7	6	2
to	3	1	5	687	3	1	7	212
eat	1	1	3	1	17	3	43	1
chinese	2	1	1	1	1	83	2	1
food	16	1	16	1	2	5	1	1
lunch	3	1	1	1	1	2	1	1
spend	2	1	2	1	1	1	1	1

Лапласын тэгшлэлттэй биграм

$$P^*(w_n|w_{n-1}) = \frac{C(w_{n-1}w_n)+1}{C(w_{n-1})+V}$$

	i	want	to	eat	chinese	food	lunch	spend
i	0.0015	0.21	0.00025	0.0025	0.00025	0.00025	0.00025	0.00075
want	0.0013	0.00042	0.26	0.00084	0.0029	0.0029	0.0025	0.00084
to	0.00078	0.00026	0.0013	0.18	0.00078	0.00026	0.0018	0.055
eat	0.00046	0.00046	0.0014	0.00046	0.0078	0.0014	0.02	0.00046
chinese	0.0012	0.00062	0.00062	0.00062	0.00062	0.052	0.0012	0.00062
food	0.0063	0.00039	0.0063	0.00039	0.00079	0.002	0.00039	0.00039
lunch	0.0017	0.00056	0.00056	0.00056	0.00056	0.0011	0.00056	0.00056
spend	0.0012	0.00058	0.0012	0.00058	0.00058	0.00058	0.00058	0.00058

Давтамжийн тоог дахин сэргээвэл

$$c^*(w_{n-1}w_n) = \frac{[C(w_{n-1}w_n)+1] \times C(w_{n-1})}{C(w_{n-1})+V}$$

	i	want	to	eat	chinese	food	lunch	spend
i	3.8	527	0.64	6.4	0.64	0.64	0.64	1.9
want	1.2	0.39	238	0.78	2.7	2.7	2.3	0.78
to	1.9	0.63	3.1	430	1.9	0.63	4.4	133
eat	0.34	0.34	1	0.34	5.8	1	15	0.34
chinese	0.2	0.098	0.098	0.098	0.098	8.2	0.2	0.098
food	6.9	0.43	6.9	0.43	0.86	2.2	0.43	0.43
lunch	0.57	0.19	0.19	0.19	0.19	0.38	0.19	0.19
spend	0.32	0.16	0.32	0.16	0.16	0.16	0.16	0.16

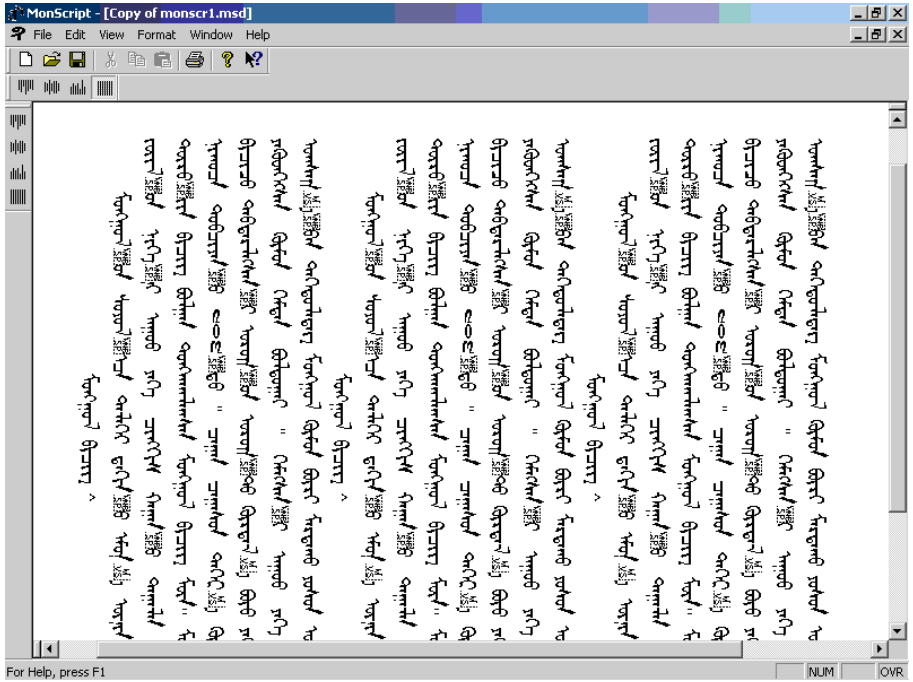
Боловсруулаагүй биграм тоотой харьцуулах

	i	want	to	eat	chinese	food	lunch	spend
i	5	827	0	9	0	0	0	2
want	2	0	608	1	6	6	5	1
to	2	0	4	686	2	0	6	211
eat	0	0	2	0	16	2	42	0
chinese	1	0	0	0	0	82	1	0
food	15	0	15	0	1	4	0	0
lunch	2	0	0	0	0	1	0	0
spend	1	0	1	0	0	0	0	0

	i	want	to	eat	chinese	food	lunch	spend
i	3.8	527	0.64	6.4	0.64	0.64	0.64	1.9
want	1.2	0.39	238	0.78	2.7	2.7	2.3	0.78
to	1.9	0.63	3.1	430	1.9	0.63	4.4	133
eat	0.34	0.34	1	0.34	5.8	1	15	0.34
chinese	0.2	0.098	0.098	0.098	0.098	8.2	0.2	0.098
food	6.9	0.43	6.9	0.43	0.86	2.2	0.43	0.43
lunch	0.57	0.19	0.19	0.19	0.19	0.38	0.19	0.19
spend	0.32	0.16	0.32	0.16	0.16	0.16	0.16	0.16

Нэгийг нэмэх үнэлгээ бүдүүлэг хэрэгсэл

- Иймээс үүнийг N-грамд хэрэглэдэггүй:
  - Илүү дээр арга олох хэрэгтэй
- Гэвч нэгийг нэмэх бусад ЭХБ-д ашиглагддаг
  - Текст ангилахад
  - Тэгийн тоо маш их байгаа бодлогод.



Хэлний загварчлал

Интерполяци, Буцах шилжилт, ба Веб-хэмжээт Х3

Буцаж шилжих ба интерполяци

- Заримдаа энэ нь **бага** агуулга ашиглахад тусалдаг
  - Сайн суралцаагүй агуулгын хувьд
- **Буцаж шилжих:**
  - Хэрэв баталгаатай бол триграм ашигла,
  - Үгүй бол биграм, эсвэл юниграм
- **Интерполяци:**
  - Юниграм, биграм, триграмыг холих
- Интерполяци практикт сайн ажилладаг.

Шугаман интерполяци

- Энгийн интерполяци

$$\hat{P}(w_n|w_{n-1}w_{n-2}) = \lambda_1 P(w_n|w_{n-1}w_{n-2}) + \lambda_2 P(w_n|w_{n-1}) + \lambda_3 P(w_n)$$

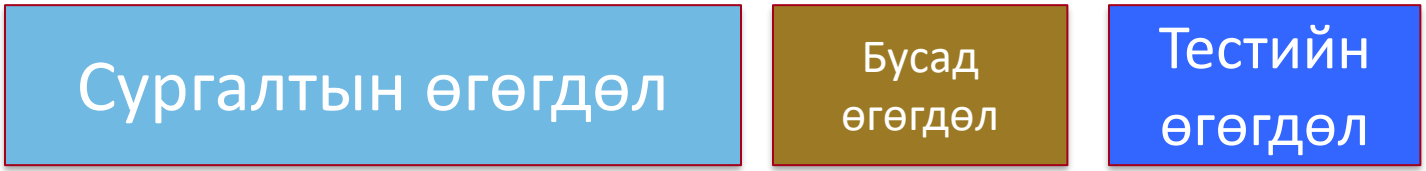
$$\sum_i \lambda_i = 1$$

- Лямбда нь орчноос хамаарна:

$$\hat{P}(w_n|w_{n-2}w_{n-1}) = \lambda_1(w_{n-2}^{n-1})P(w_n|w_{n-2}w_{n-1}) + \lambda_2(w_{n-2}^{n-1})P(w_n|w_{n-1}) + \lambda_3(w_{n-2}^{n-1})P(w_n)$$

Лямбдаг хэрхэн тодорхойлох вэ?

- **бусад үгсийн корпус** ашиглах



- Бусад өгөгдлийн магадлалыг хамгийн их байлгахын тулд λ-г сонгох:
  - N-грам магадлалыг засах (сургалтын өгөгдөл дээр)
  - Дараа нь бусад үгсийн олонлогт хамгийн их магадлал өгөх λ-г хайх: $\log P(w_1 \ldots w_n \mid M(l_1 \ldots l_k)) = \sum_i \log P_{M(l_1 \ldots l_k)}(w_i \mid w_{i-1})$

Мэдэхгүй үгс: Нээлттэйн эсрэг хаалттай үгийн сангийн бодлогууд

- Хэрэв бүх үгийг мэддэг бол
  - Үгийн сан V бол тогтмол
  - Хаалттай үгийн сангийн бодлого
- Гэвч бид бүх үгийг үргэлж мэдээд байдаггүй
  - **Үгийн сангаас гаднах** = ҮСГ үгс /Out Of Vocabulary – OOV /
  - Нээлттэй үгийн сангийн бодлого
- Оронд нь: мэдэхгүй үгийн токен үүсгэ <UNK>
  - <UNK> магадлалын сургалт
    - V үгийн сангаас L тогтмол үгийн сан үүсгэнэ
    - Текст нормчлох шатанд L –д байхгүй сургалтын үг бүрийг <UNK> болгож өөрчлөнө
    - Эцэст нь түүний магадлалыг энгийн үг шиг сургана
  - Код тайлах үед
    - Хэрэв текст оролт: сургаагүй дурын үгийн хувьд UNK магадлалыг ашиглана.

Том веб-хэмжээт n-грам

- Хэрхэн шийдвэрлэх вэ, ж.нь, Google N-грам корпус
- Тайралт - Pruning
  - Зөвхөн хязгаараас давсан тоотой N-грамуудыг хадгалах.
    - Өндөр зэрэглэлийн n-грамаас хос бишүүдийг хасах
  - Энтропид суурилсан тайралт
- Бүтээмж сайжруулах
  - Мод зэрэг үр дүнтэй өгөгдлийн бүтэц
  - Блуумын шүүр: хэлний загварыг ойролцоолох
  - Индекс байдлаар үгийг хадгалах, холбоосоор биш
    - Олон үгийг багасгаж 2 байт болгохын тулд Хаффмены кодыг ашиглана.
  - Магадлалыг тоо хэлбэрт шилжүүлэх (8-н байтын хөвөгч таслалтай тооны оронд 4-8 бит)

Веб-хэмжээт N-грамыг тэгшлэх

- “Ухаангүй буцаж шилжих” алгоритм (Brants *et al*. 2007)
- Бууруулахгүй, зөвхөн харьцангуй давтамж ашиглана.

$$S(w_i \mid w_{i-k+1}^{i-1}) = \begin{cases} \dfrac{\text{count}(w_{i-k+1}^{i})}{\text{count}(w_{i-k+1}^{i-1})} & \text{if } \text{count}(w_{i-k+1}^{i}) > 0 \\ 0.4S(w_i \mid w_{i-k+2}^{i-1}) & \text{otherwise} \end{cases}$$

$$S(w_i) = \frac{\text{count}(w_i)}{N}$$

N-грам тэгшлэлтийн дүгнэлт

- Нэгийг нэмэх тэгшлэлт:
  - Текст ангилалтанд асуудалгүй, хэлний загварчлалд тохирохгүй
- Хамгийн их өргөн ашигладаг арга:
  - Кнессер-Нейн өргөтгөсөн интерполяци
- Веб шиг маш том N-грамын хувьд:
  - Ухаангүй буцаж шилжилт

Хэлний дэвшилтэт загварчлал

- Дискреминант/ялгавартай/ загварууд:
  - Сургалтын олонлогыг тохируулахад биш, харин үр дүнг сайжруулахын тулд n-грамын жинг сонгох
- Задлан шинжилгээнд суурилсан загвар
- Кейшлэх загвар
  - Саяхан ашигласан үгс ахин гарч ирэх магадлалтай

$$P_{CACHE}(w \mid history) = / P(w_i \mid w_{i-2} w_{i-1}) + (1 - /) \frac{c(w \hat{I} \; history)}{|history|}$$

  - Эдгээр нь яриа танилтанд маш бага магадлалтай байдаг. Яагаад?