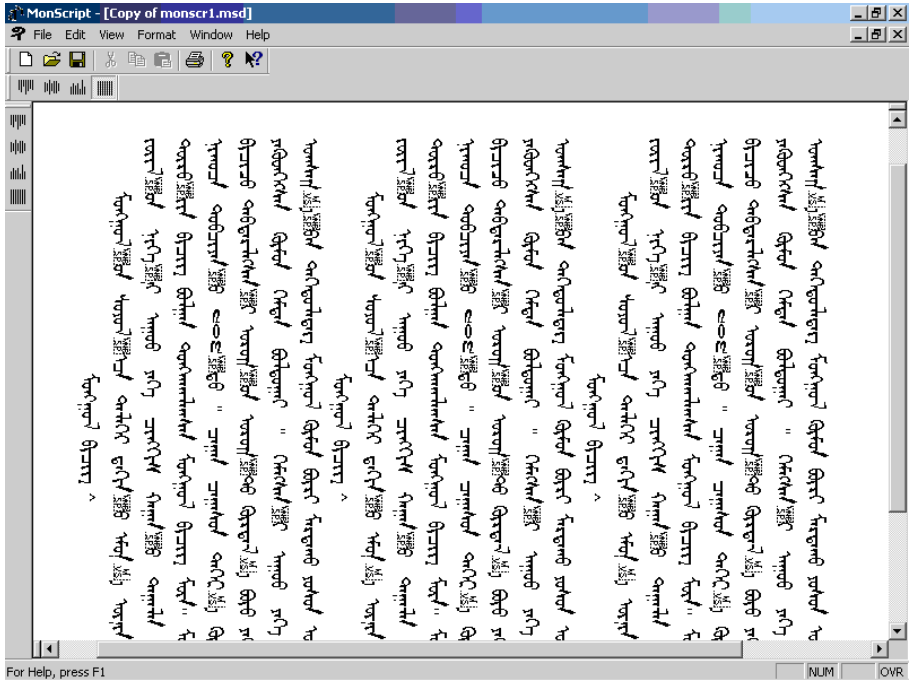


Хэлний загварчлал

Лекц №7

Н-грамын тухай

2020 он

Хэлний магадлалын загвар

- Өнөөдрийн зорилго: өгүүлбэрийн магадлал олох

Яагаад?

- Машин орчуулга:
  - P(**high** winds tonite) > P(**large** winds tonite)
- Зөв бичгийн алдааг зүгшрүүлэлт
  - The office is about fifteen **minuets** from my house
    - P(about fifteen **minutes** from) > P(about fifteen **minuets** from)
- Яриа танилт
  - P(I saw a van) >> P(eyes awe of an)
- + Дүгнэлт, асуултанд хариулах, гэх мэт!!

Хэлний магадлалын загвар

- Зорилго: өгүүлбэр эсвэл үгсийн дарааллын магадлал бодох:

$$P(W) = P(w_1,w_2,w_3,w_4,w_5 \ldots w_n)$$

- Хамааралтай бодлого: тухайн үгийн магадлал:

$P(w_5|w_1,w_2,w_3,w_4)$ $P(w_1,w_2,w_3,w_4,w_5)$

- Эдгээрийн аль алиныг тооцоолдог загвар:

$P(W)$ эсвэл $P(w_n|w_1,w_2 \ldots w_{n-1})$ -г **хэлний загвар** гэж хэлдэг.

- Илүү: **дүрэм** Гэвч **хэлний загвар** эсвэл **ХЗ** гэх нь стандарт

P(W) –г хэрхэн тооцоолох вэ?

- Нийлмэл магадлалыг хэрхэн тооцоолох вэ:
  - P(its, water, is, so, transparent, that)
- Төсөөлөл: Магадлалын гинжин дүрэмд найдацгаая

Санамж: Гинжин дүрэм

- Нөхцөлт магадлалын тодорхойлолтыг эргэж санавал

$P(A|B)=P(A,B)/P(B)$ Rewriting: $P(A|B)\,P(B) = P(A,B)$

$P(A,B) = P(A|B)\,P(B)$

- Илүү олон хувьсагч:

$P(A,B,C,D) = P(A)P(B|A)P(C|A,B)P(D|A,B,C)$

- Ерөнхий тохиолдолд гинжин дүрэм

$$P(x_1,x_2,x_3,\ldots,x_n) = P(x_1)P(x_2|x_1)P(x_3|x_1,x_2)\ldots P(x_n|x_1,\ldots,x_{n-1})$$

Гинжин дүрмийг өгүүлбэр дэх үгсийн нийлмэл магадлалын тооцоолоход хэрэглэдэг.

$$P(w_1 w_2 \cdots w_n) = \prod_i P(w_i \mid w_1 w_2 \cdots w_{i-1})$$

P(“its water is so transparent”) =

P(its) × P(water|its) × P(is|its water)

× P(so|its water is) × P(transparent|its water is so)

Эдгээр магадлалыг хэрхэн үнэлэх вэ?

- Зөвхөн тоолоод, хуваахад болох уу?

$$P(\text{the} \mid \text{its water is so transparent that}) = \frac{Count(\text{its water is so transparent that the})}{Count(\text{its water is so transparent that})}$$

- ! Хэтэрхий олон боломжит өгүүлбэрүүд байна!
- Эдгээрийг үнэлэхэд хэрэгтэй хангалттай өгөгдөл хэзээ ч олдохгүй

Марковын таамаглал


Andrei Markov

- Таамаглалыг хялбарчилбал:

$$P(\text{the} \mid \text{its water is so transparent that}) \approx P(\text{the} \mid \text{that})$$

- Эсвэл магадгүй

$$P(\text{the} \mid \text{its water is so transparent that}) \approx P(\text{the} \mid \text{transparent that})$$

Марковын таамаглал

$$P(w_1 w_2 \cdots w_n) \approx \prod_i P(w_i \mid w_{i-k} \cdots w_{i-1})$$

- Өөрөөр хэлбэл, үржигдэхүүн бүрийг ойролцоолно

$$P(w_i \mid w_1 w_2 \cdots w_{i-1}) \approx P(w_i \mid w_{i-k} \cdots w_{i-1})$$

Хамгийн энгийн тохиолдол: Юниграм загвар

$$P(w_1 w_2 \ldots w_n) \approx \prod_i P(w_i)$$

Юниграм загвараас үүссэн зарим автоматаар үүссэн өгүүлбэрүүд

fifth, an, of, futures, the, an, incorporated, a, a, the, inflation, most, dollars, quarter, in, is, mass

thrift, did, eighty, said, hard, 'm, july, bullish

that, or, limited, the

Биграм загвар

- Өмнөх үгэнд нөхцөлдүүлбэл:

$$P(w_i \mid w_1 w_2 \cdots \; w_{i-1}) \approx P(w_i \mid w_{i-1})$$

texaco, rose, one, in, this, issue, is, pursuing, growth, in, a, boiler, house, said, mr., gurria, mexico, 's, motion, control, proposal, without, permission, from, five, hundred, fifty, five, yen

outside, new, car, parking, lot, of, the, agreement, reached

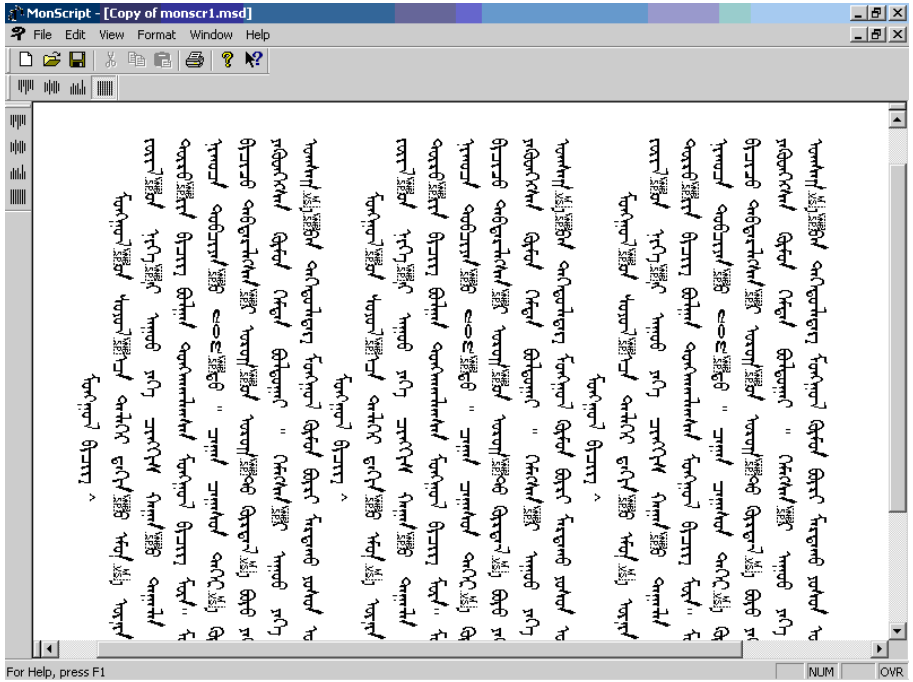
this, would, be, a, record, november

N-грам загварууд

- триграм, 4-грам, 5-грам гээд өргөжүүлж болно
- Ерөнхийдөө энэ бол хэлний хувьд хангалтгүй загвар
  - Учир нь хэл нь **хол-зайн хамааралтай** байдаг:

  “Миний компьютер өчигдөр зүгээр ажиллаж байгаад өнөөдөр гацчихлаа.”

- Гэхдээ N-грам загвар асуудлыг шийдэх боловч төгс шийдэл биш



Хэлний загварчлал

N-грам магадлалыг үнэлэх нь

Биграм магадлалыг үнэлэх

- Хамгийн их үнэний хувийн үнэлгээ

$$P(w_i \mid w_{i-1}) = \frac{count(w_{i-1}, w_i)}{count(w_{i-1})}$$

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

Жишээ 1

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

<s> I am Sam </s>
<s> Sam I am </s>
<s> I do not like green eggs and ham </s>

$P(\texttt{I} \mid \texttt{<s>}) = \frac{2}{3} = .67$ $P(\texttt{Sam} \mid \texttt{<s>}) = \frac{1}{3} = .33$ $P(\texttt{am} \mid \texttt{I}) = \frac{2}{3} = .67$

$P(\texttt{</s>} \mid \texttt{Sam}) = \frac{1}{2} = 0.5$ $P(\texttt{Sam} \mid \texttt{am}) = \frac{1}{2} = .5$ $P(\texttt{do} \mid \texttt{I}) = \frac{1}{3} = .33$

Жишээ 2: Ресторантай холбоотой өгүүлбэрүүд

- can you tell me about any good cantonese restaurants close by
- mid priced thai food is what i’m looking for
- tell me about chez panisse
- can you give me a listing of the kinds of food that are available
- i’m looking for a good place to eat breakfast
- when is caffe venezia open during the day

Боловсруулаагүй биграм тоолуур

- 9222 өгүүлбэрээс авсан

	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

Боловсруулаагүй биграм магадлал

- Юниграмаар нормчилбол:

i	want	to	eat	chinese	food	lunch	spend
2533	927	2417	746	158	1093	341	278

- Үр дүн:

	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

Өгүүлбэрийн магадлалуудын биграм үнэлгээ

P(<s> I want english food </s>) =

P(I|<s>)

× P(want|I)

× P(english|want)

× P(food|english)

× P(</s>|food)

= .000031

Мэдлэгийн ямар төрлүүд байна?

- P(english|want) = .0011
- P(chinese|want) = .0065
- P(to|want) = .66 - infinitive дүрэм
- P(eat | to) = .28
- P(food | to) = 0 - гэнэтийн тэг
- P(want | spend) = 0 - дүрмийн, бүтцийн тэг
- P (i | <s>) = .25

Практик хүндрэлүүд

- Бүгдийг лог хэлбэрээр хийдэг
  - Утга алдагдахаас зайлсхий
  - (мөн нэмэх нь үржүүлэхээс хурдан)

$$\log(p_1 \cdot p_2 \cdot p_3 \cdot p_4) = \log p_1 + \log p_2 + \log p_3 + \log p_4$$

Хэлний загварчлалын хэрэгсэлүүд

- SRILM
  - http://www.speech.sri.com/projects/srilm/

Google N-Грам хувилбар, 2006 8 сар

AUG
3

All Our N-gram are Belong to You

Posted by Alex Franz and Thorsten Brants, Google Machine Translation Team

Here at Google Research we have been using word n-gram models for a variety of R&D projects,

...

That's why we decided to share this enormous dataset with everyone. We processed 1,024,908,267,229 words of running text and are publishing the counts for all 1,176,470,663 five-word sequences that appear at least 40 times. There are 13,588,391 unique words, after discarding words that appear less than 200 times.

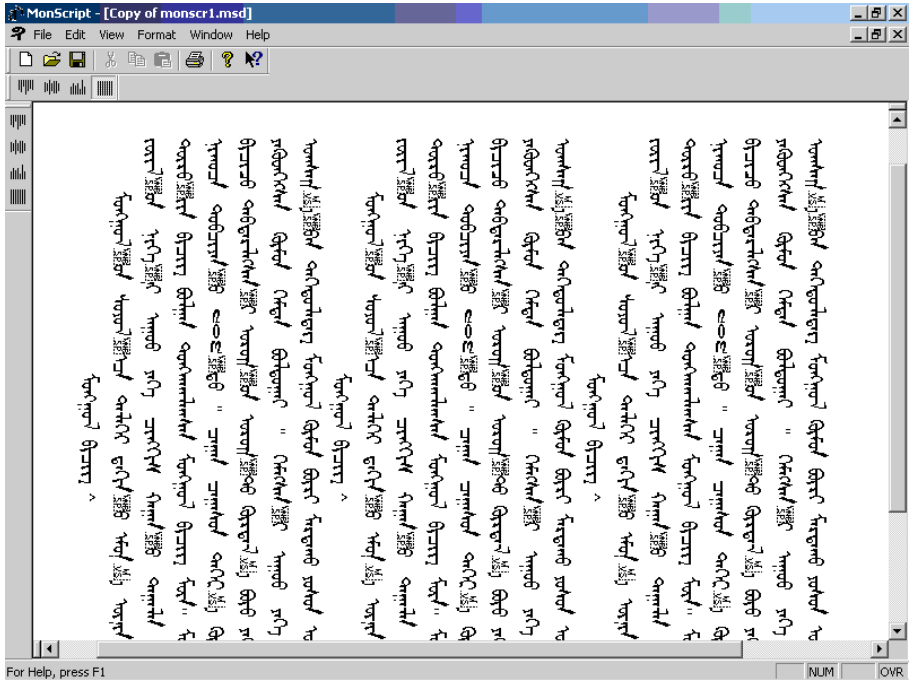
Google N-Грам хувилбар

- serve as the incoming 92
- serve as the incubator 99
- serve as the independent 794
- serve as the index 223
- serve as the indication 72
- serve as the indicator 120
- serve as the indicators 45
- serve as the indispensable 111
- serve as the indispensible 40
- serve as the individual 234

http://googleresearch.blogspot.com/2006/08/all-our-n-gram-are-belong-to-you.html

Google N-грам ном

- https://books.google.com/ngrams



Хэлний загварчлал

Үнэлгээ ба эргэлзээт чадвар

Үнэлгээ: Загвар чинь хэр сайн байна?

- Хэлний загвар чинь зөв өгүүлбэрийг буруугаас илүүд үзэж байна уу?
  - Өндөр магадлалыг “бодит” эсвэл “байнга тохиолдох” өгүүлбэрт өгөх
    - Мөн “дүрмийн бус” эсвэл “цөөхөн тохиолдох” өгүүлбэрт?
- Загварын параметруудыг **сургалтын олонлог** дээр сургадаг.
- Загварын гүйцэтгэлийг ашиглаагүй өгөгдөл дээр шалгадаг.
  - **Тестийн олонлог** нь огт ашиглаагүй, сургалтын олонлогоос ялгаатай өгөгдлийн олонлог байна.
  - **Үнэлгээний хэмжигдэхүүн** нь загвар тестийн олонлогт хэр сайн ажиллахыг харуулна.

N-грам загварын хөндлөнгийн үнэлгээ (extrinsic)

- А ба В загваруудыг харьцуулах хамгийн сайн үнэлгээ
  - Загвар бүрт нэг даалгавар өгнө
    - Зөв бичгийн алдаа засагч, яриа танигч, машин орчуулгын систем
  - Даалгаврын ажиллуулж, А болон В –ийн зөв утгын нарийвчлалыг авна
    - Хэдэн буруу үгийг зөв зассан
    - Хэдэн үгийг зөв орчуулсан
  - А ба В –н зөв утгын нарийвчлалыг харьцуул

N-грам загварын хөндлөнгийн үнэлгээний (in-vivo) хүндрэл

- Хөндлөнгийн үнэлгээ
  - Цаг үрдэг; хэдэн өдөр, хэдэн 7 хоног болдог. Удаан ажилладаг.
- Иймээс
  - Заримдаа дотоод(intrinsic) үнэлгээг ашигладаг: **эргэлзээт чадвар** (нийтлэг дотоод үнэлгээ - perplexity)
  - Энэ нь тестийн өгөгдөл нь сургалтын өгөгдөлтэй хэтэрхий төстэй байвал хөндлөнгийн үнэлгээнээс муу үр дүн үзүүлэх боломжтой.
    - Иймээс **зөвхөн урьдчилсан туршилтанд ашигтай**
  - Гэхдээ хөндлөнгийн үнэлгээнээс их ашигладаг.

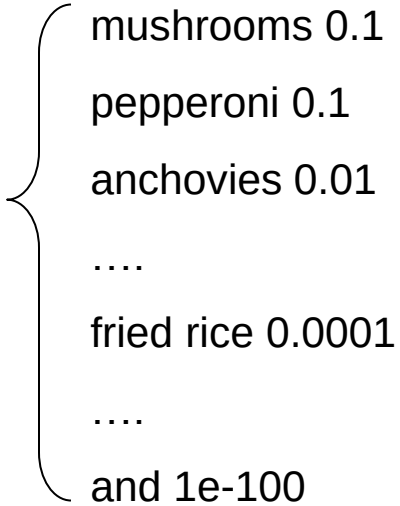
Эргэлзээт чадварын төсөөлөл

- Шаноны тоглоом:
  - Дараагийн үгийг хэрхэн зөв таамаглах вэ?

    I always order pizza with cheese and ____

    The 33rd President of the US was ____

    I saw a ____



  - Юниграм энэ тоглоомонд тохирохгүй. (Яагаад?)
- Хэлний зөв загвар нь
  - үнэхээр тохиолддог үгэнд өндөр магадлал олдог.

Эргэлзээт чадвар

Хамгийн сайн хэлний загвар нь өмнө оруулаагүй тестийн олонлог дээр сайн таамагладаг.

- Хамгийн өндөр P(өгүүлбэр)

Эргэлзээт чадвар нь үгсийн тоогоор нормчлогдсон, тестийн олонлогын урвуу магадлал:

$$PP(W) = P(w_1 w_2 ... w_N)^{-\frac{1}{N}}$$

$$= \sqrt[N]{\frac{1}{P(w_1 w_2 ... w_N)}}$$

Гинжин дүрэм:

$$\mathrm{PP}(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i | w_1 \ldots w_{i-1})}}$$

Биграмын хувьд:

$$\mathrm{PP}(W) = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{P(w_i | w_{i-1})}}$$

Хамгийн бага эргэлзээт чадвар нь хамгийн их магадлалтай ижил

Шаноны тоглоомын эргэлзээт чадварын төсөөлөл

- Josh Goodman зохиосон салаалалтын дундаж нөхцөл
- ‘0,1,2,3,4,5,6,7,8,9’ цифр таних даалгавар хэр хэцүү вэ
  - эргэлзээт чадвар 10
- Microsoft –д (30,000) нэр таних хэр хэцүү вэ.
  - эргэлзээт чадвар = 30,000
- Эргэлзээт чадвар бол ижил салаалалтын нөхцөл

Эргэлзээт чадвар салаалах нөхцөл

- Өгүүлбэр нь санамсаргүй цифрүүдээс бүрдэнэ гэж үзье
- Цифр бүрд P=1/10 ноогддог загварын хувьд энэ өгүүлбэрийн эргэлзээт чадвар хэд вэ?

$$
\begin{aligned}
\mathrm{PP}(W) &= P(w_1 w_2 \ldots w_N)^{-\frac{1}{N}} \\
&= \left(\frac{1}{10}^{N}\right)^{-\frac{1}{N}} \\
&= \frac{1}{10}^{-1} \\
&= 10
\end{aligned}
$$

Бага эргэлзээт чадвар = сайн загвар

- 38 сая үгээр сургаж, 1.5 сая үгээр шалгав, Wall Street Journal

N-грамын дараалал	Юниграм	Биграм	Триграм
эргэлзээт чадвар	962	170	109