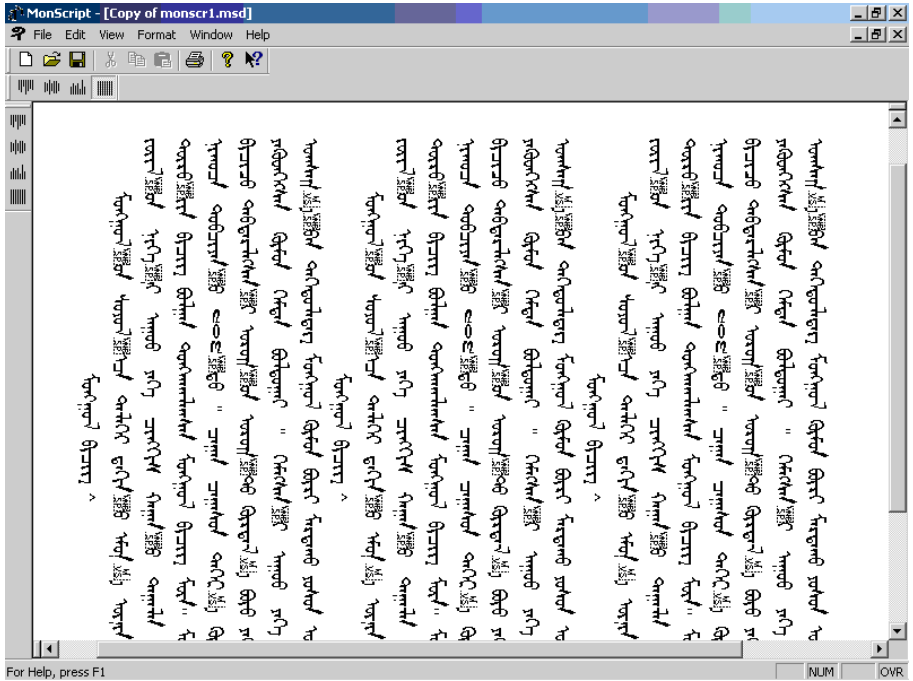
Үгийн алдаа зүгшрүүлэлт ба шуугиант суваг



Зөв бичиглэлтэй үгийг засах

Зөв бичиглэлтэй үгийн алдаа

- …leaving in about fifteen ***minuets*** to go to her house.
- The design ***an*** construction of the system…
- Can they ***lave*** him my messages?
- The study was conducted mainly ***be*** John Black.

- зөв бичгийн алдааны 25-40% нь зөв бичиглэлтэй үг байдаг Kukich 1992

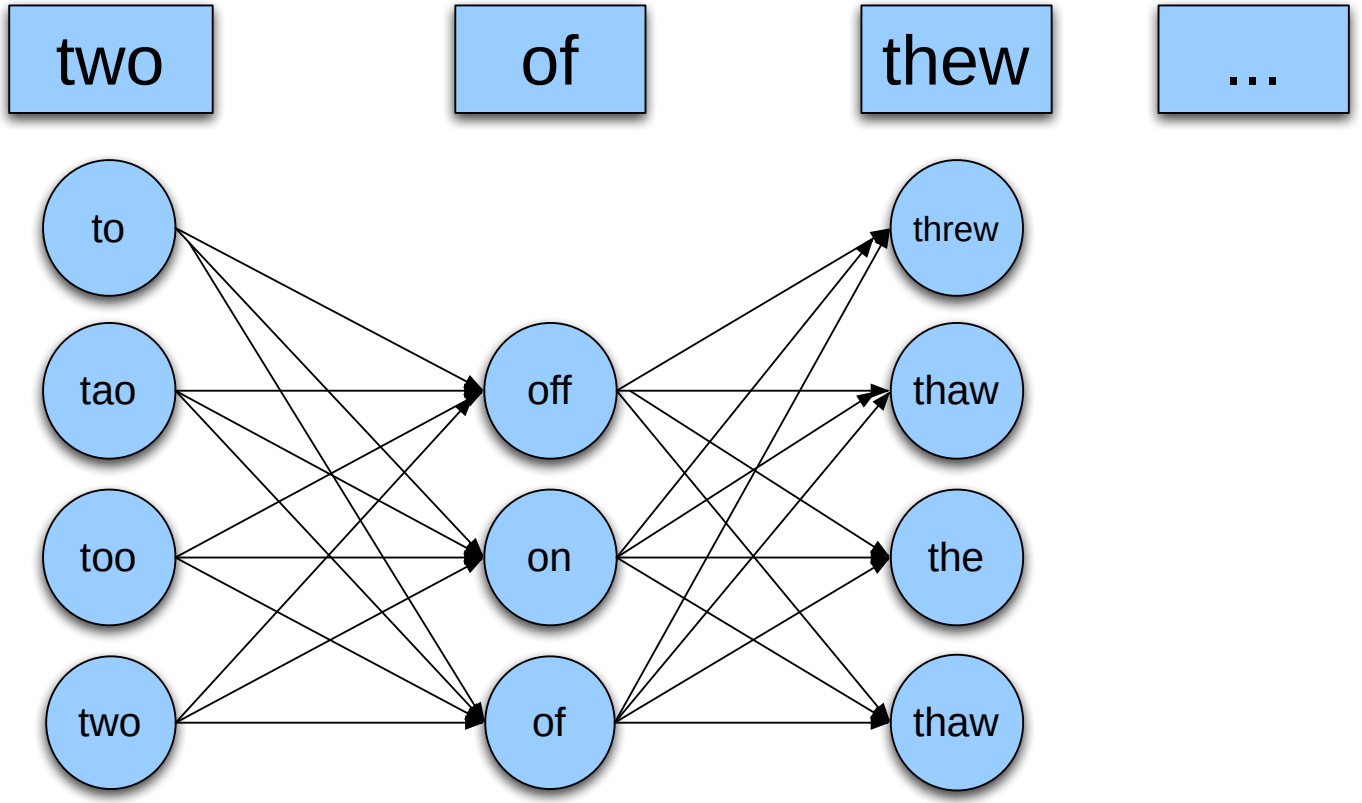
Зөв бичиглэлтэй үгийн алдааг шийдэх нь

- Өгүүлбэр доторх үг бүрийн хувьд
  - *Санал болгох үгийн олонлогийг* үүсгэ
    - тухайн үг өөрөө
    - нэг үсгийн зөрүүтэй бүх англи үгс
    - ижил дуудлагатай үгс
- Хамгийн сайн тохирох үгсийг сонгох
    - Шуугиант сувгийн загвар
    - Даалгавар заасан ангилагч

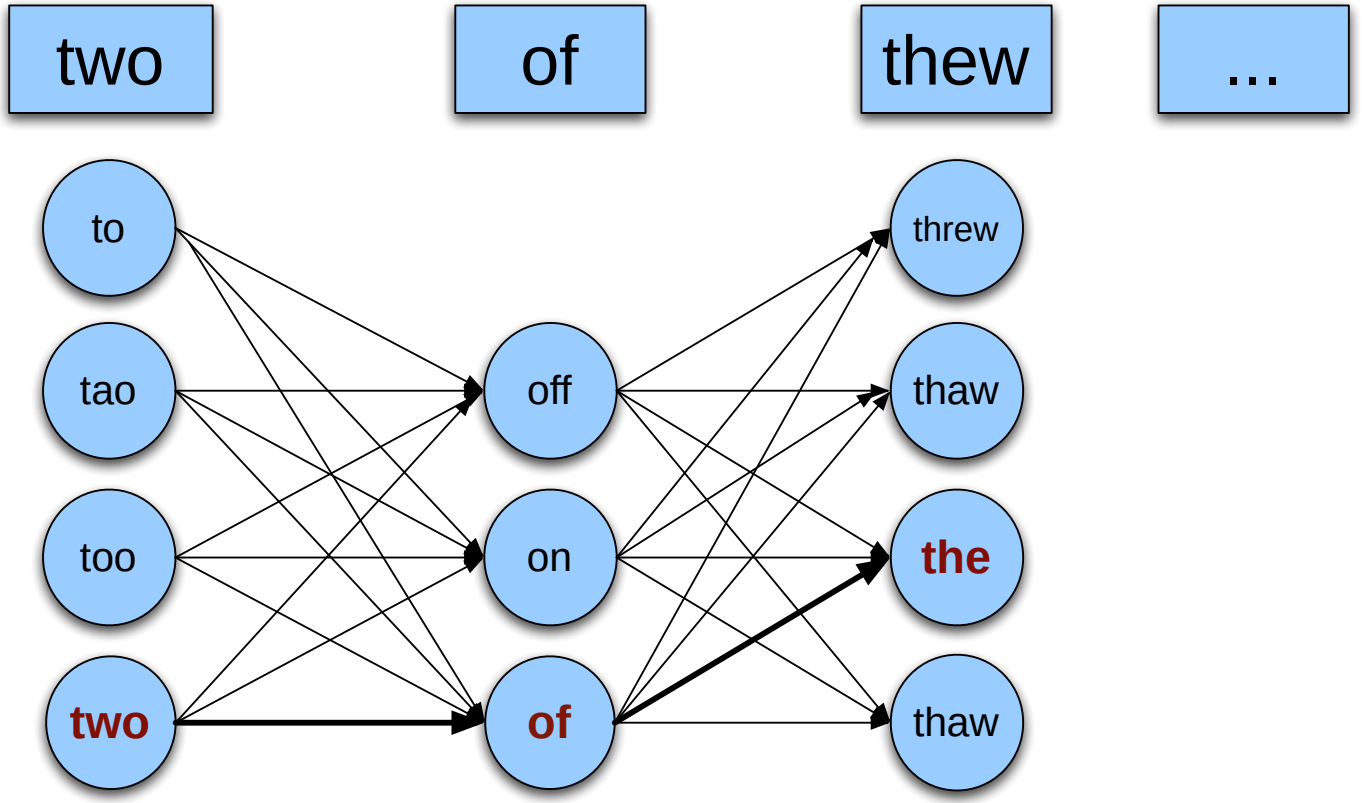
Зөв бичиглэлтэй үгийн алдааг зүгшрүүлэлтийн шуугиант суваг

- $w_1, w_2, w_3, \ldots, w_n$ үгс бүхий өгөгдсөн нэг өгүүлбэр
- w_i үг бүрийн хувьд санал болгох үгсийн олонлог үүсгэ
  - Candidate(w_1) = {w_1, w'_1 , w''_1 , w'''_1 ,…}
  - Candidate(w_2) = {w_2, w'_2 , w''_2 , w'''_2 ,…}
  - Candidate(w_n) = {w_n, w'_n , w''_n , w'''_n ,…}
- P(W) магадлалыг хамгийн их байлгах W дарааллыг сонго

Зөв бичиглэлтэй үгийн алдааг зүгшрүүлэлтийн шуугиант суваг



Зөв бичиглэлтэй үгийн алдааг зүгшрүүлэлтийн шуугиант суваг



Хялбарчлал: өгүүлбэр бүрт нэг алдаа

- Нэг үг солигдсон бүх боломжит өгүүлбэрүүдийг гарга
  - w_1, **w’’**$_2$,w_3,w_4 two **off** thew
  - w_1,w_2,**w’**$_3$,w_4 two of **the**
  - **w’’’**$_1$,w_2,w_3,w_4 **too** of thew
  - …
- P(W) магадлалыг хамгийн их байлгах W дарааллыг сонго

Магадлалуудыг хаанаас авах вэ

- Хэлний загвар
  - Юниграм
  - Биграм
  - гэх мэт.
- Сувгийн загвар
  - Буруу бичигдсэн үгийн зүгшрүүлэлттэй ижил
  - Харин бусад үгс буруу бичигдсэн байж болох P(w|w) алдааны бус магадлал нэмэгддэг.

Алдааны бус магадлалыг хэрхэн тооцоолох

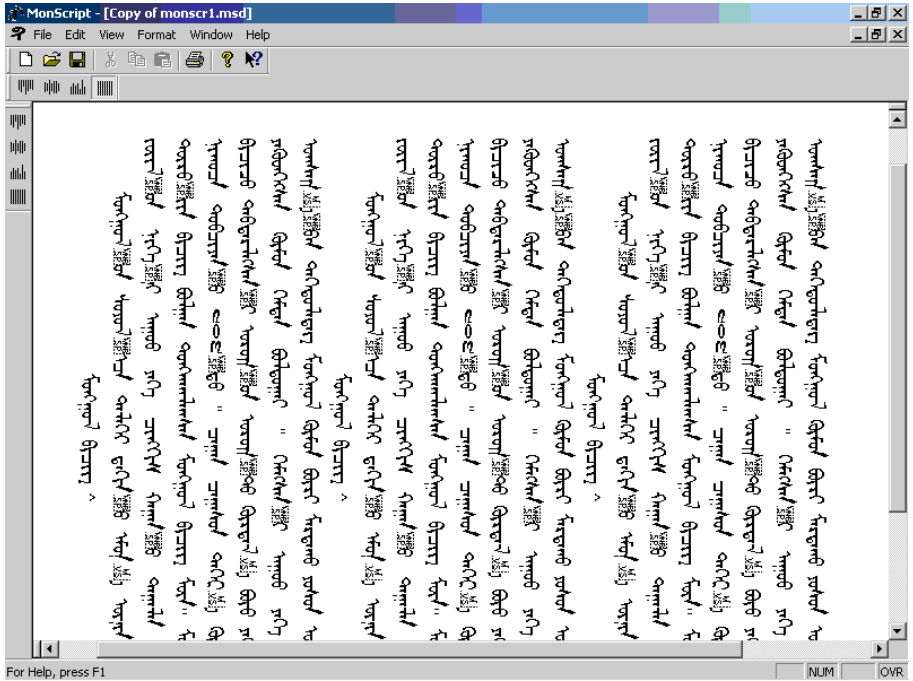
- Зөв бичсэн үгийн хувьд сувгийн магадлал ямар байх вэ?
- P(“the”|“the”)

- мэдээж энэ нь тухайн програмаас хамаарна
  - .90 (10 үгэнд 1 алдаа)
  - .95 (20 үгэнд 1 алдаа)
  - .99 (100 үгэнд 1 алдаа)
  - .995 (200 үгэнд 1 алдаа)

Питер Норвигийн “thew” жишээ

x	w	x|w	P(x|w)	P(w)	10^9 P(x|w)P(w)
thew	the	ew|e	0.000007	0.02	144
thew	thew		0.95	0.00000009	90
thew	thaw	e|a	0.001	0.0000007	0.7
thew	threw	h|hr	0.000008	0.000004	0.03
thew	thwe	ew|we	0.000003	0.00000004	0.0001

Үгийн алдаа зүгшрүүлэлт ба шуугиант суваг



Орчин үеийн системүүд

зөв бичгийн алдаа шалгалт дахь HCI асуудлууд

- Хэрэв зүгшрүүлэлтэд их итгэлтэй байвал
  - автомат зүгшрүүлэлт
- Бага итгэлтэй бол
  - хамгийн сайн тохирох зүгшрүүлэлтийг өгөх
- Арай бага итгэлтэй байвал
  - зүгшрүүлэлтийн жагсаалтыг өгөх
- Илтгэлгүй байвал
  - зөвхөн алдааны флаг

HCI – human computer interaction – хүн компьютерийн харилцаа

Орчин үеийн шуугиант суваг

- Бид зөвхөн өмнөх болон алдааны загвар дахь магадлалыг хэзээ ч үржүүлдэггүй
- Хараат бус таамаглал →магадлалуудыг харьцуулашгүй
- Оронд нь: Тэднийг жигнэ

$$\hat{w} = \underset{w \hat{\mathrm{I}} V}{\arg\max} P(x \mid w) P(w)^{l}$$

- Хөгжүүлэлтийн тестийн олонлогоос λ –г сурга

Авиа зүйн алдааны загвар

- GNU aspell –д ашигладаг мета-авиа (metaphone)
  - зөв бичгийн алдааг мета-авиа дуудлага руу хөрвүүлэх
    - “C –ээс бусад, зэргэлдээ үсгүүдийн давхцалыг арилга”
    - “Хэрэв үг 'KN', 'GN', 'PN', 'AE', 'WR’ –ээр эхэлсэн бол, эхний үсгийг арилга”
    - “'M’ –н дараах болон үгийн төгсгөлийн ‘B’-г арилга”
    - …
  - алдаатай үгээс 1-2 засварын хэмжээтэй үгсийг ол
  - үр дүнгийн жагсаалтыг дүгнэх
    - Буруу бичсэн үг болон санал болгох үг хоорондын засварын жигнэсэн хэмжээ
    - Буруу дуудлага болон санал болгох дуудлага хоорондын засварын хэмжээ

Сувгийн загварыг сайжруулах

- Өөр бусад засварыг зөвшөөрөх (Brill and Moore 2000)
  - ent→ant
  - ph→f
  - le→al
- Дуудлагын загварыг сувгийн загварттай нэгтгэх (Toutanova and Moore 2002)

Сувгийн загвар

- p(алдаатай үг|үг) магадлалд нөлөөлж болох хүчин зүйлс
  - эх үсэг
  - зорилгын үсэг
  - эргэн тойрны үсгүүд
  - үгэн дэх байрлал
  - компьютерийн гаран дээрх ойр орчмын товчнууд
  - компьютерийн гаран дээрх төстэй товчнууд
  - дуудлагууд
  - бүтээвэр хувирах магадлал

Хөрш үсгүүд

Зөв бичиглэлтэй үгийн зүгшрүүлэлтийн хувьд ангилагчид суурилсан аргууд

- Зөвхөн сувгийн болон хэлний загварын оронд
- нэг ангилагч доторх олон онцлог шинжүүдийг ашигла (дараагийн лекц дэх).
- тухайн нэг хосын хувьд нэг ангилагч үүсгэ:

  whether/weather

  - +- 10 үгс доторх “cloudy”
  - ___ to VERB
  - ___ or not