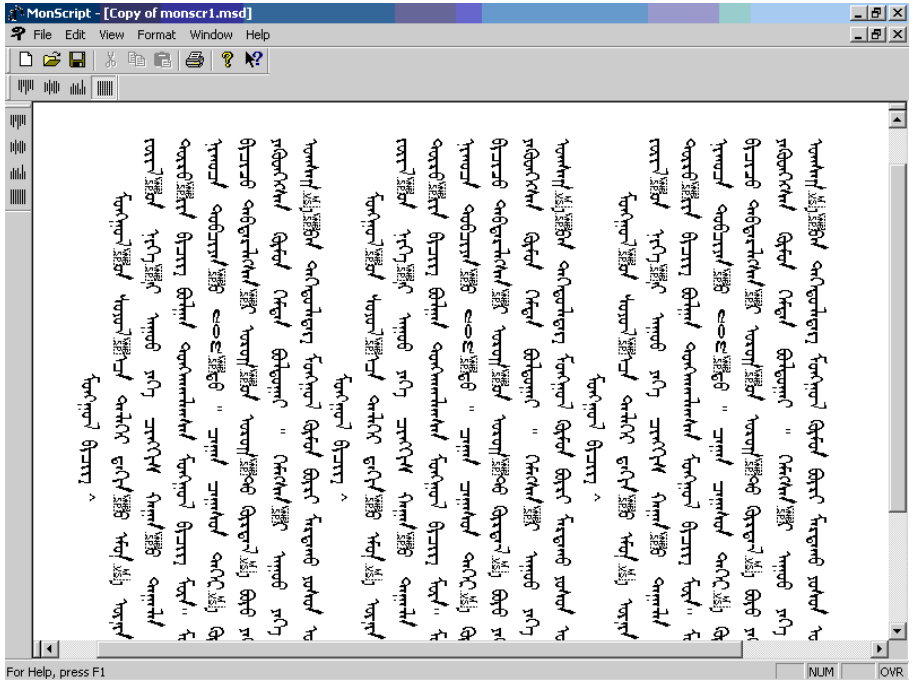


Эх хэлний боловсруулалтын тухай ойлголт

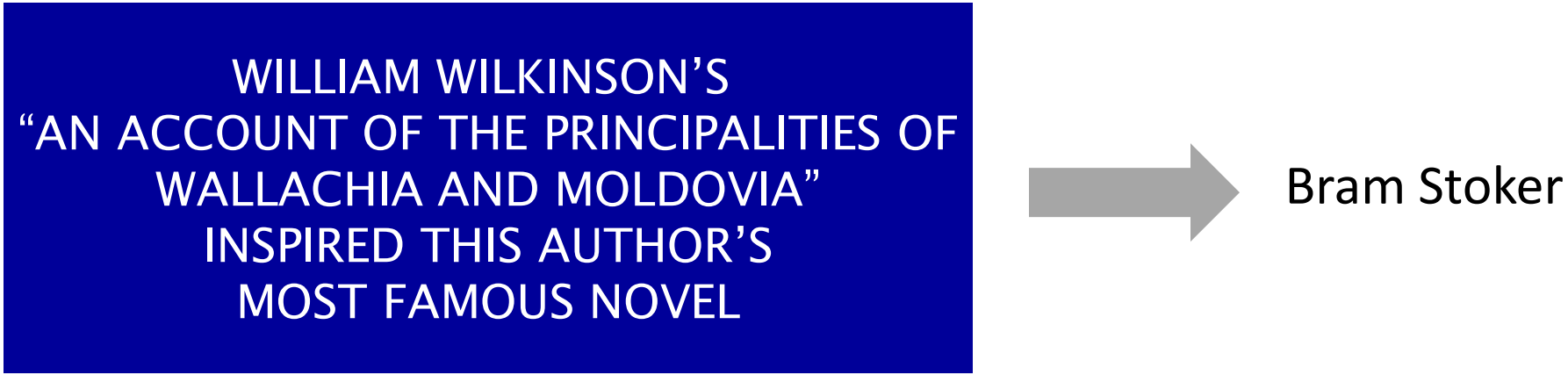
Лекц №1

Эх хэлний боловсруулалт гэж юу вэ ?

2020 он

Асуултанд хариулах (Question answering)

- 2011 онд IBM’s Watson буюу Jeopardy тэмцээнд ийм системээр ялж байсан.



Мэдээлэл гаргах (Information Extraction)

Subject: **curriculum meeting**

Date: January 15, 2012

To: Dan Jura

Event: Curriculum mtg
Date: Jan-16-2012
Start: 10:00am
End: 11:30am
Where: Gates 159

Hi Dan, we've now scheduled the curriculum meeting.

It will be in Gates 159 tomorrow from 10:00-11:30.

-Chris

Мэйлээс календарт бүртгэл оруулна.

Мэдээлэл гаргаж, хандлагийн (sentiment) шинжилгээ

Шинж чанарууд:

томруулалт

чадамж

хэмжээ болон жин

гэрэл

хэрэглэхэд хялбар

Нийлбэр дүн:

шүүмжлэл:

Хэмжээ болон жин

- ✓ Авч явахад амар, цомхон
- ✓ Жижигхэн хөнгөн камер байхад хүнд, мэргэжлийн камер авч явах хэрэггүй.
- ✗ Нимгэн гэж мэдрэгдэх камер нь хуванцар бөгөөд хэт хөнгөн тиймээс ийм камерыг удирдахад их хэцүү байдаг.

Машин орчуулга

- Бүрэн автомат

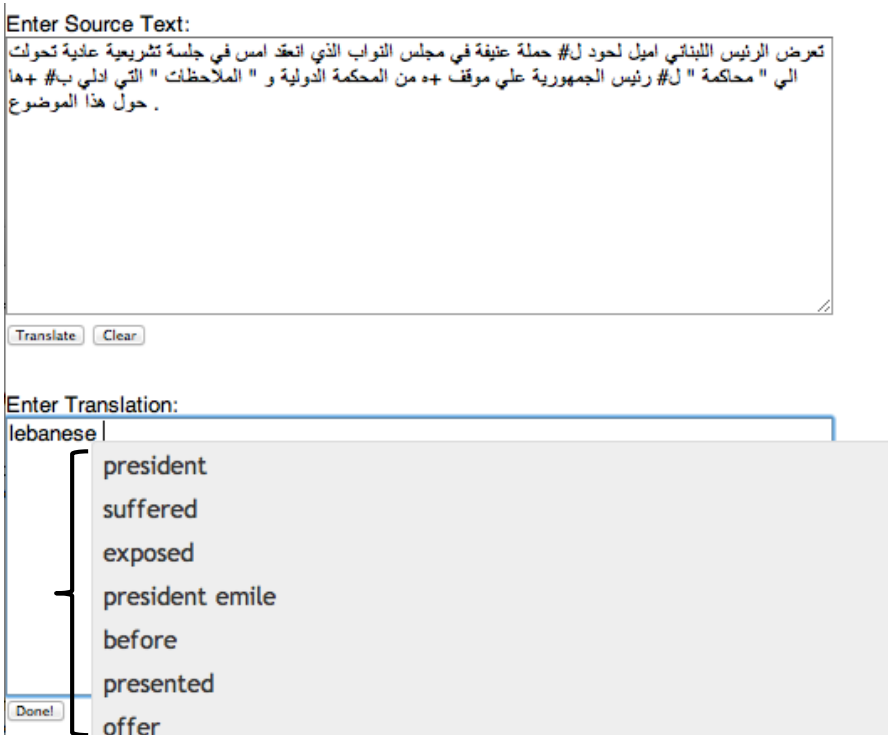
Эх текст оруул:

这 不过 是 一 个 时间 的 问题 .

Станфорд-ийн хэллэгээр орчуулбал:

This is only a matter of time.

- Орчуулагчид туслах



Хэлний технологи

Ихэнхдээ шийдсэн байдаг

Спам таних

Let's go to Agra! ✓

Buy V1AGRA … ✗

Үгийн аймаг таних (POS)

ADJ	ADJ	NOUN	VERB	ADV
Colorless	green	ideas	sleep	furiously.

Оноосон нэр үг таних (NER)

PERSON	ORG	LOC

Einstein met with UN officials in Princeton

Явц сайтай байгаа

Хандлагийн шинжилгээ

Best roast chicken in San Francisco!

The waiter ignored us for 20 minutes.

Нэг зүйлийг олох-Coreference resolution

Carter told Mubarak he shouldn't run again.

Үгийн утга таних (WSD)

I need new batteries for my ***mouse***.

Өгүүлбэрзүйн шинжилгээ (Parsing)

I can see Alcatraz from the window!

Машин орчуулга (MT)

第13届上海国际电影节开幕…

The 13th Shanghai International Film Festival…

Мэдээлэл гаргах (IE)

You're invited to our dinner party, Friday May 27 at 8:30

Party
May 27
add

Одоо хүртэл шийдэгдээгүй

Асуултанд хариулах (QA)

Q. How effective is ibuprofen in reducing fever in patients with acute febrile illness?

Найруулга хийх (Paraphrase)

XYZ acquired ABC yesterday

ABC has been taken over by XYZ

Дүгнэх

The Dow Jones is up

The S&P500 jumped

Housing prices rose

⇨ Economy is good

Харилцан яриа

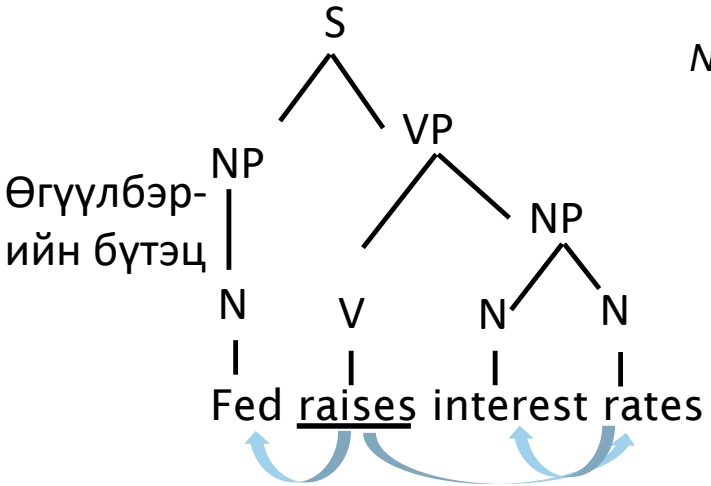
Where is Citizen Kane playing in SF?

Castro Theatre at 7:30. Do you want a ticket?

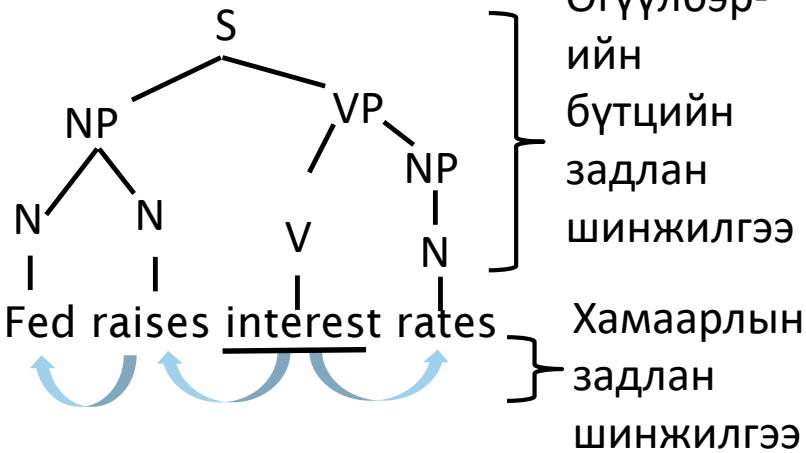
Олон салаа утгатай, тодорхойгүй байдал нь ЭХБ-ыг хүндрүүлдэг.

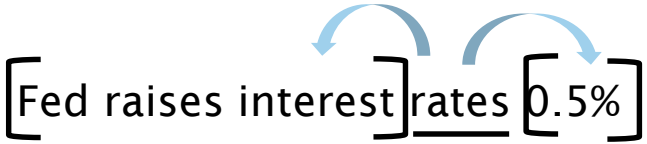
- Утгын хувьд олон утгатай үг
  - “сур” үгийн утга
    - Сурах (үйл)
    - Эрж сурах (үйл)
    - Сур харвах (нэр)
- Үгийн бүтцээр задлахад олон хэлбэрээр задрах
  - “өглөө” үгийн задралт
    - өг+лөө (“өг” үгэнд өнгөрсөн цагийн “лөө” залгах)
    - өглөө (өглөө, тэмдэг нэр)

Салаа утгатай байдал хэлэнд их тохиолддог.



New York Times headline (17 May 2000)





Яагаад эх хэлийг ойлгох хэцүү вэ?

Стандарт бус бичиглэл

Great job @justinbieber! Were SOO PROUD of what youve accomplished! U taught us 2 #neversaynever & you yourself should never give up either♥

Сегментлэх асуудал

in-law

the New York-New Haven Railroad

the New York-New Haven Railroad

Хэлц үг - idioms

dark horse
get cold feet
lose face
throw in the towel

Шинэ үг бүтээх neologisms

unfriend
Retweet
bromance

world knowledge

Mary and Sue are sisters.

Mary and Sue are mothers.

Түвэгтэй оноосон нэр

Where is *A Bug's Life* playing …

Let It Be was recorded …

… a mutation on the *for* gene …

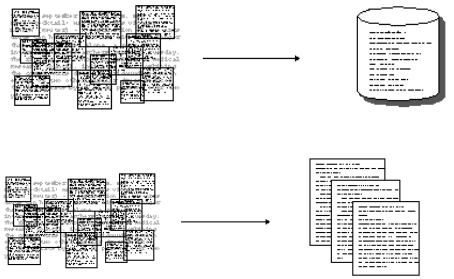
Гэвч сонирхолтой

Энэ асуудал дээр ахиц гаргах нь...

- Бодлогын даалгавар хүнд! Ямар хэрэгслүүд бидэнд хэрэгтэй вэ?
  - Хэлний тухай мэдлэг
  - Ертөнцийн тухай мэдлэг
  - Мэдлэгийн эх үүсвэрүүдийг хослуулах арга зам
- Үүний бид ерөнхий хэрхэн хийх вэ гэвэл:
  - Хэлний өгөгдлөөс магадлалын модел байгуулна
    - P(“maison” → “house”) өндөр магадлал
    - P(“L’avocat général” → “the general avocado”) бага магадлал
  - Боловсруулаагүй текстийн шинж чанарууд нь ажлын хагасыг нугалсан байдаг.

Энэ хичээлээр

- Статистик ЭХБ-ын онол, аргууд:
  - Маркавын далд загвар
  - Naïve Bayes, максимум интропи ангилагчууд
  - N-gram хэлний модел
  - Статистик шинжилгээ
  - Урвуу индекс, утгын вектор модел
- Лабораторид бодит амьдрал дээрх хэрэглээ:
  - Мэдээлэл гаргах
  - Зөв бичгийн алдаа зүгшрүүлэлт
  - Мэдээлэл хайх
  - Хандлагийн шинжилгээ



Шаардлагатай ур чадварууд

- Энгийн шугаман алгебр (вектор, матрици)
- Магадлалын суурь онол
- Java эсвэл Python програмчлалын хэл