## CS-AD 220 – Spring 2016

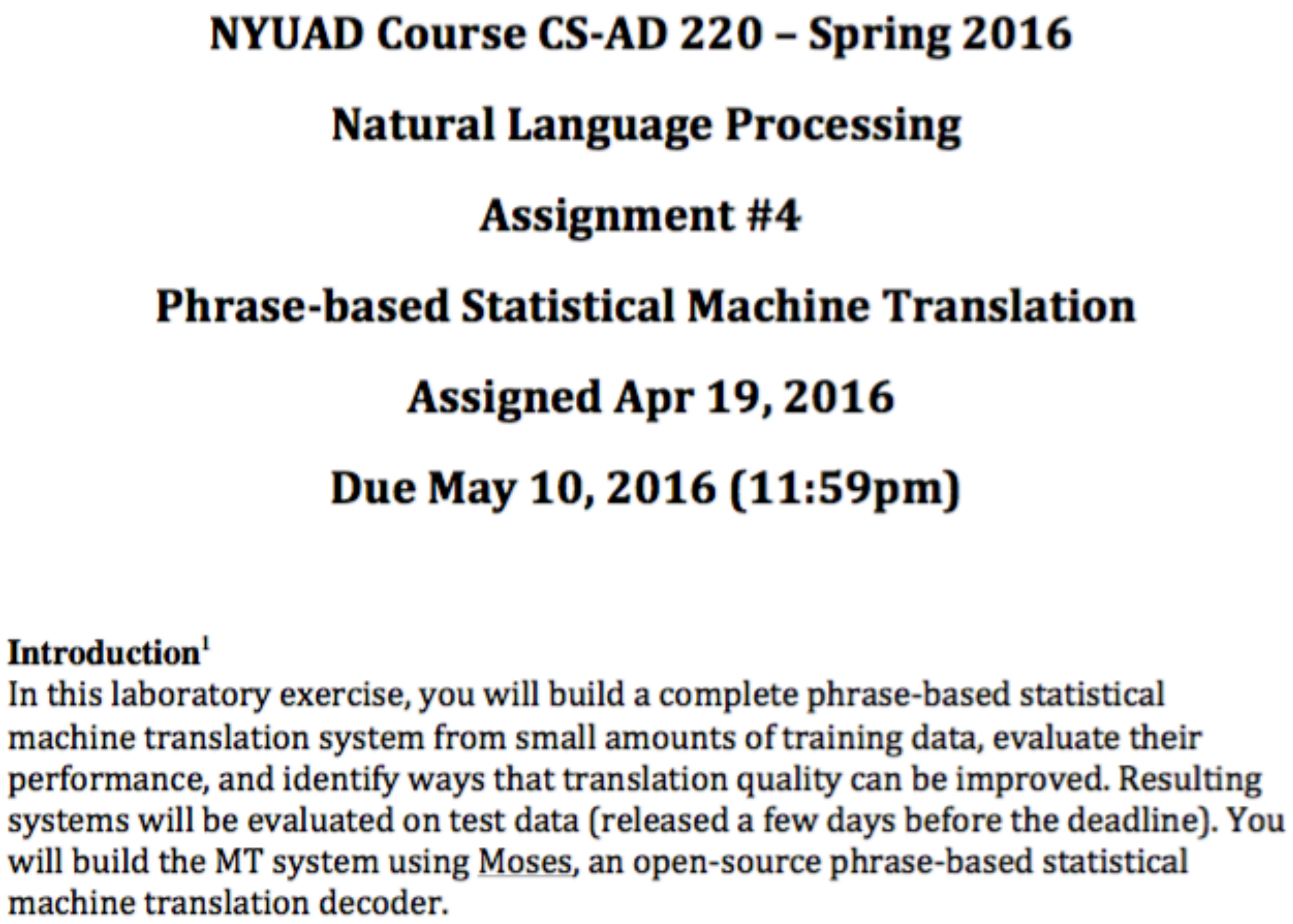
# Natural Language Processing

### Session 22: 19-Apr-16

Prof. Nizar Habash & Prof. Hieu Hoang

# Looking ahead

- Hackathon Results!
- Deadline Assignment #3
  - April 17 → April 22
- Prof. Jan Hajic
  - April 21 (room ERB 120)



# NYUAD Course CS-AD 220 – Spring 2016

# Natural Language Processing

# Assignment #4

# Phrase-based Statistical Machine Translation

## Assigned Apr 19, 2016

## Due May 10, 2016 (11:59pm)

**Introduction**[1]

In this laboratory exercise, you will build a complete phrase-based statistical machine translation system from small amounts of training data, evaluate their performance, and identify ways that translation quality can be improved. Resulting systems will be evaluated on test data (released a few days before the deadline). You will build the MT system using Moses, an open-source phrase-based statistical machine translation decoder.

*Assignment #4 posted on NYU Classes*

*START EARLY!*

*DEADLINE IS May 10 (11:59pm)*

# Assignment 4

Class Meet Moses!

# Start

- Get the USB key with the MT Assignment virtual machine from Hieu or Nizar

- Computer with +30GB disk space

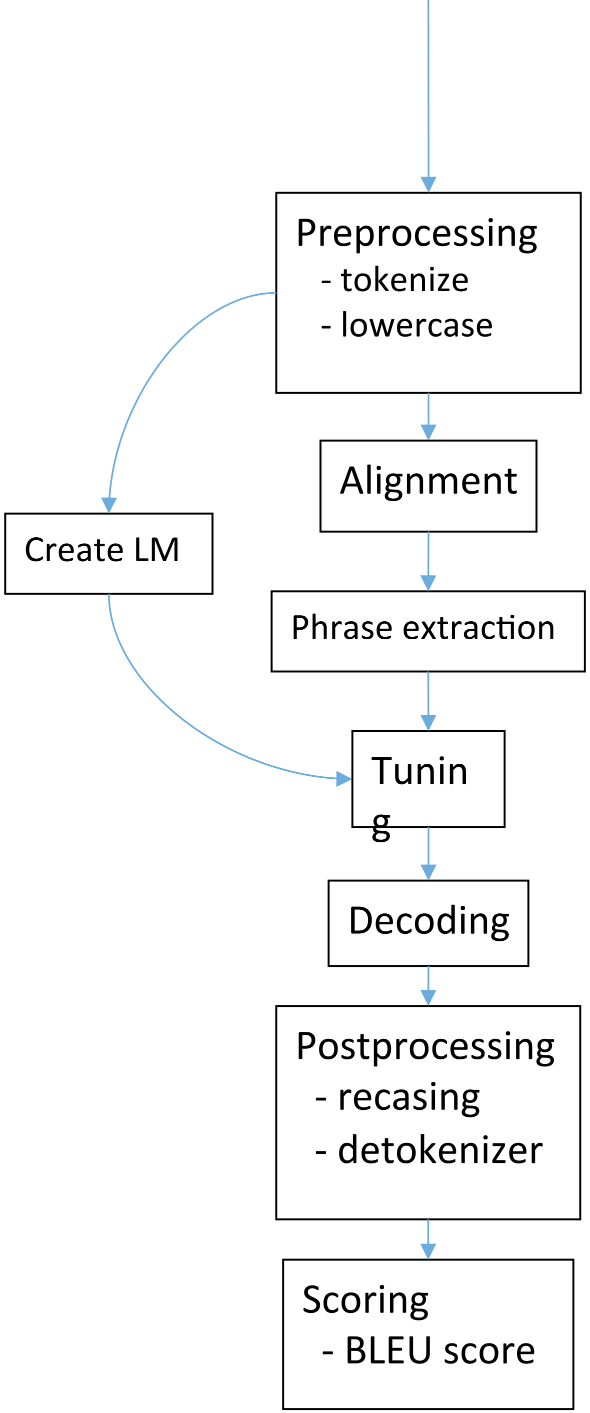
- Follow the instructions!
  - Install VirtualBox
  - Run virtual machine (Ubuntu Linux)
  - Run commands

- Creating Arabic-to-English translation system

# Data

- Arabic – Buckwalter encoding (’Romanized’)
  - AlOx gyr Alcqyq lSdAm Hsyn yrfD AlEwdp IlY AlErAq
- Datasets
  - Train
    - 35,644 parallel sentences
    - 71,286 sentences just in English
  - Tune
    - 50 parallel sentences
  - Test
    - 48 parallel sentences

# SMT Pipeline



Preprocessing
- tokenize
- lowercase

Alignment

Create LM

Phrase extraction

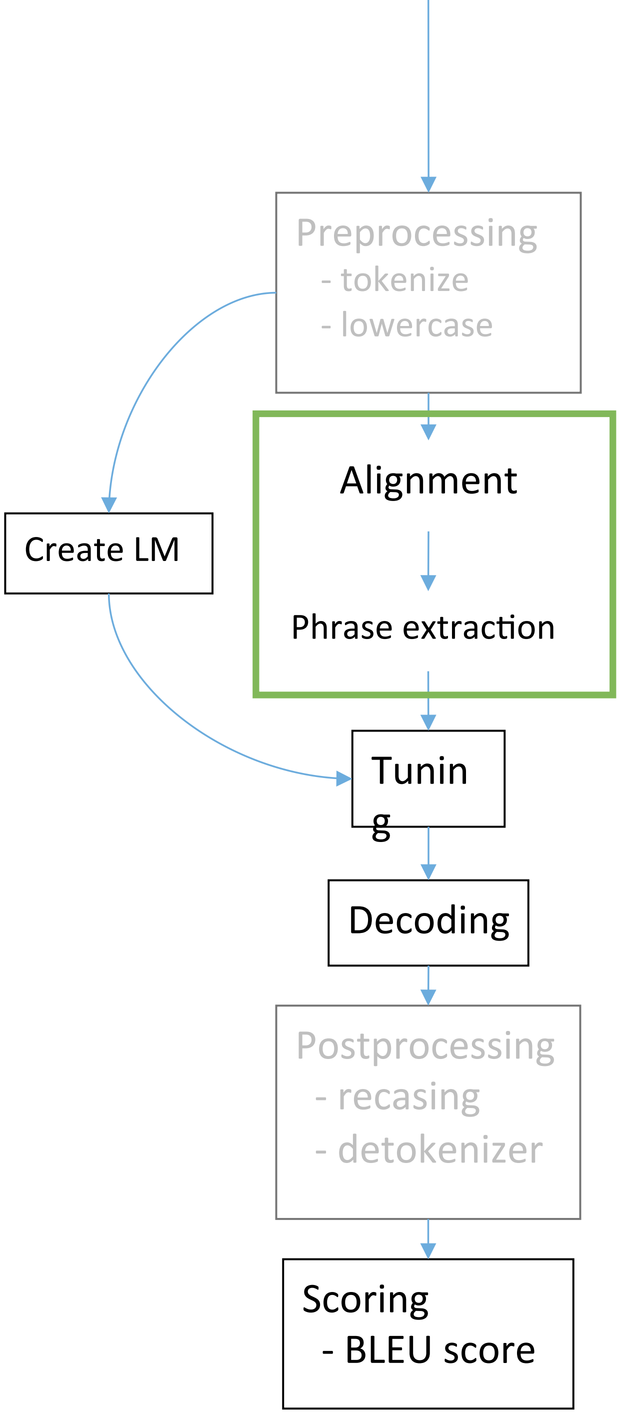
Tunin
g

Decoding

Postprocessing
- recasing
- detokenizer

Scoring
- BLEU score

# SMT Pipeline



Preprocessing
- tokenize
- lowercase

Create LM

Alignment

Phrase extraction

Tunin
g

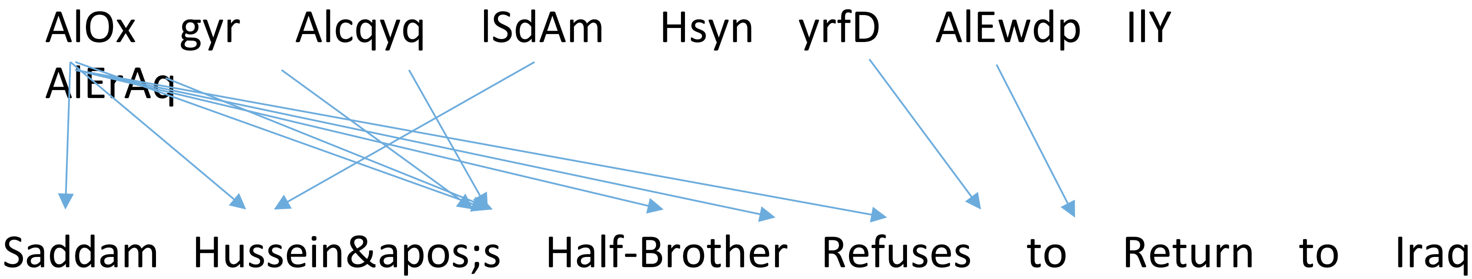
Decoding

Postprocessing
- recasing
- detokenizer

Scoring
- BLEU score

# Word Alignment

- Training data
  - data/Train/Train_data.clean.[en/ar]
- Word alignment
  - work/model/aligned.grow-diag-final-and
  - Eg. 0-0 0-1 4-1 0-2 1-2 2-2 3-2 0-3 0-4 0-5 7-6 8-7



AlOx gyr Alcqyq lSdAm Hsyn yrfD AlEwdp IlY AlErAq

Saddam Hussein&apos;s Half-Brother Refuses to Return to Iraq

# Phrase-Table

! ! ! . . ||| **People pass by houses** ||| 0.2 5.34133e-10 0.166667 4.38429e-14 ||| 0-1 ||| 5 6 1 |||

**source** **target** **p(s|t)** **p(t|s)**

- 3.7 million translation rules
  - 73MB zipped, 422MB unzipped
  - Too slow to load all into memory
  - Use too much RAM

- Filter phrase table
  - Only keep rules need to translate the test set

# Language Model

Target text: **the cow jumped over the moon**

**p(the cow jumped over the moon)** =

p(the) *
p(cow|the) *
p(jumped| the cow) *
p(over| the cow jumped) *
p(the|the cow jumped over) *
p(moon| the cow jumped over the)

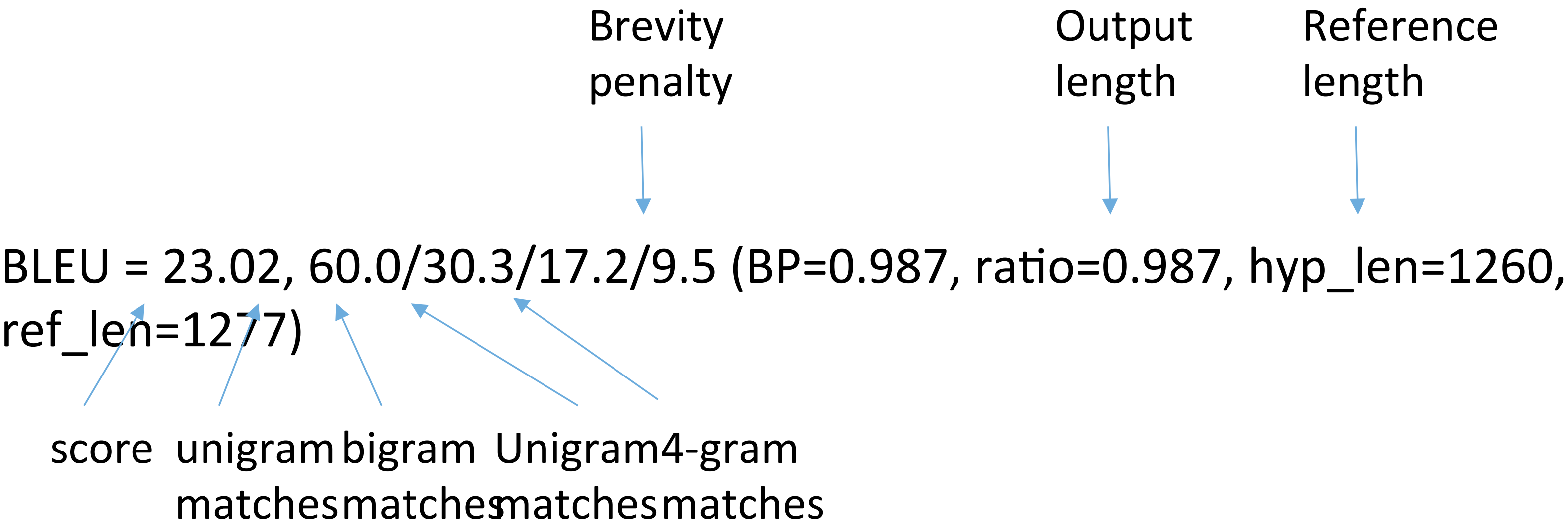
≈ p(the) *
p(cow|the) *
p(jumped| the cow) *
p(over| ~~the~~ cow jumped) *
p(the|~~the cow~~ jumped over) *
p(moon| ~~the cow jumped~~ over the)

File **work/LM/LM_data+Train_data.en.lm**

```
\data\
ngram 1=139572
ngram 2=1061731
ngram 3=2239731

\1-grams:
-6.0734353    <unk>     0
0             <s>       -0.91558355
-1.6365006    </s>      0
-5.2046447    Nicosia   -0.11571049
....
\2-grams:
-2.1021864    (AFP) </s>  0
-1.4692371    - </s>      0
....
\3-grams:
-0.16613887    <s> (AFP) </s>
-1.4355018    18/02 (AFP) </s>
....
```

# BLEU score



Brevity penalty

Output length

Reference length

```
BLEU = 23.02, 60.0/30.3/17.2/9.5 (BP=0.987, ratio=0.987, hyp_len=1260, ref_len=1277)
```

score

unigram matches

bigram matches

Unigram matches

4-gram matches

# Sınav #5

# Quiz (in Turkish) #5

# Quiz #5

Names:____________________

# A Trip to Istanbul!

Adapted from *Come to Istanbul* (NACLO 2014)

- Turkish is spoken by about 63 million people, most of whom live in Turkey. Turkish is a non-Indo-European language, so it is unrelated to English but related to languages of Central Asia such as Azeri and Uzbek.
- Turkish words are built up by adding one or more endings to a root word; the vowels in most word endings vary depending on the vowels in the root word ("vowel harmony"), as you will see in the following examples. Here are some sentences in Turkish, with their English translations. Note:
    - *The Turkish letters "ş", "ç" and "ı" are pronounced like English "sh", "ch" and the "a" in "above".*
    - *The letters i and ı represent different vowels.*
    - *The letter "ğ" is usually silent (like the "gh" in "although").*
    - *Square brackets [ ] enclose English words that are not directly translated.*
- Use the provided examples to learn enough Turkish words and rules to translate these three sentences.

| | |
|---|---|
| İstanbul en büyük şehir<br>Istanbul [is the] biggest city. | Arkadaşlarım şehirde mutlu<br>My friends [are] happy in [the] city. |
| Eve geliyorlar<br>They come home. | Baban İstanbul'u seviyor mu?<br>Does your father like Istanbul? |
| Pencereden atlıyoruz<br>We jump from [the] window. | Evimizde büyük pencereler var<br>There are big windows in our house. |
| Fakirler Van'dan İstanbul'a gelmek istiyor<br>Poor [people] want to come from Van to Istanbul. | Ev almak mı istiyorsun?<br>Do you want to buy [a] house? |
| Babam "Merhaba! Gel, arkadaşımız ol", diyor<br>My father says "Hello! Come [and] be our friend". | |

Baban mutlu mu?

____________________________________

Arkadaşım doktor olmak istiyor.

____________________________________

İstanbul'dan mı geliyorsun?

____________________________________

# Turkish-English Parallel Text

| | |
|---|---|
| Arkadaşlarım şehirde mutlu | My friends [are] happy in [the] city. |
| Baban İstanbul'u seviyor mu? | Does your father like Istanbul? |
| Fakirler Van'dan İstanbul'a gelmek istiyor | Poor [people] want to come from [the city of] Van to Istanbul. |
| İstanbul en büyük şehir | Istanbul [is the] biggest city. |
| Eve geliyorlar | They come home. |
| Babam "Merhaba! Gel, arkadaşımız ol", diyor | My father says "Hello! Come [and] be our friend”. |
| Evimizde büyük pencereler var | There are big windows in our house. |
| Pencereden atlıyoruz | We jump from [the] window. |
| Ev almak mı istiyorsun? | Do you want to buy [a] house? |

# Translate these sentences:

- Baban mutlu mu?
- Arkadaşım doktor olmak istiyor.
- İstanbul'dan mı geliyorsun?

# Quiz #5

Names:____________________

# A Trip to Istanbul!

Adapted from *Come to Istanbul* (NACLO 2014)

- Turkish is spoken by about 63 million people, most of whom live in Turkey. Turkish is a non-Indo-European language, so it is unrelated to English but related to languages of Central Asia such as Azeri and Uzbek.
- Turkish words are built up by adding one or more endings to a root word; the vowels in most word endings vary depending on the vowels in the root word ("vowel harmony"), as you will see in the following examples. Here are some sentences in Turkish, with their English translations. Note:
  - *The Turkish letters "ş", "ç" and "ı" are pronounced like English "sh", "ch" and the "a" in "above".*
  - *The letters i and ı represent different vowels.*
  - *The letter "ğ" is usually silent (like the "gh" in "although").*
  - *Square brackets [ ] enclose English words that are not directly translated.*
- Use the provided examples to learn enough Turkish words and rules to translate these three sentences.

| | |
|---|---|
| İstanbul en büyük şehir<br>Istanbul [is the] biggest city. | Arkadaşlarım şehirde mutlu<br>My friends [are] happy in [the] city. |
| Eve geliyorlar<br>They come home. | Baban İstanbul'u seviyor mu?<br>Does your father like Istanbul? |
| Pencereden atlıyoruz<br>We jump from [the] window. | Evimizde büyük pencereler var<br>There are big windows in our house. |
| Fakirler Van'dan İstanbul'a gelmek istiyor<br>Poor [people] want to come from Van to Istanbul. | Ev almak mı istiyorsun?<br>Do you want to buy [a] house? |
| Babam "Merhaba! Gel, arkadaşımız ol", diyor<br>My father says "Hello! Come [and] be our friend". | |

Baban mutlu mu?

Is your father happy?

Arkadaşım doktor olmak istiyor.

my friend wants to become a doctor.

İstanbul'dan mı geliyorsun?

do you come from Istanbul?

# Facts about Turkish

## *Agglutinatination*

| Turkish | English |
|---|---|
| ev | (the) house |
| evler | (the) houses |
| evin | your (sing.) house |
| eviniz | your (pl./formal) house |
| evim | my house |
| evimde | at my house |
| evlerinizin | of your houses |
| evlerinizden | from your houses |
| evlerinizdendi | (he/she/it) was from your houses |
| evlerinizdenmiş | (he/she/it) was (apparently/said to be) from your houses |

# Facts about Turkish
# *Nouns*

- No definite article, No gender
- X+s (plural) = X-lar or X-ler
- Possessive
  - my X (X-ım, -am), your X (X-an) our X (X–imiz)
- In X ➔ X-de
- From X ➔ X-den
- To X ➔ X-a or X-e
- Subject X ➔ Ø
- Object X ➔-u

# Facts about Turkish
## *Verbs*

- Verb tend to come at the end of sentence.
- "to V” ➔ V-mak or V-mek
- Basic V ➔ –iyor-<Subject>
  - He/She/It V (-iyor)
  - We V (-iyor-uz)
  - They V (iyor-lar)
  - You V (iyor-sun)

# Facts about Turkish
# *Vowel Harmony*

- Turkish has eight vowels
  - Front vowels: i ü e ö
  - Back vowels: ı u a o
- Words may not contain a mix of back/front

# Facts about Turkish
# *Vowel Harmony*

- Turkish has eight vowels
  - Front vowels: i ü e ö
  - Back vowels: ı u a o
- Words cannot contain both front and back vowels
  - el+ler (hand+s) ➔ eller
  - kız+ler (girl+s) ➔ kız+lar

# Next Time

- No reading
- Come with questions about the MT assignment