Introduction to Statistical Machine Translation

David Kauchak
CS159 – Spring 2011

Some slides adapted from

Philipp Koehn
CSAIL
Massachusetts Institute of Technology

Kevin Knight
USC/Information Sciences Institute
USC/Computer Science Department

Admin

- How did assignment 5 go?
- Project proposals?
  - I will give you feedback soon
- Start working on the projects!
- Quiz on Wednesday

Quiz #3

- text similarity
  - set overlap methods
  - vector-based methods
    - different distance metrics
    - weighting schemes: IDF and PMI
- word similarity
  - character-based
  - semantic web-based
  - dictionary-based
  - distributional/similarity-based
- misc topics:
  - stoplist
  - WordNet
  - edit distance
- information retrieval
  - general problems, evaluation, etc.
  - papers/student presentations

Language translation

MT Systems

Where have you seen machine translation systems?

Machine Translation

美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件，威胁将会向机场等公众地方发动生化袭击後，关岛经保持高度戒备。 ➡ The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport .

The classic acid test for natural language processing.

Requires capabilities in both interpretation and generation.

People around the world stubbornly refuse to write everything in English.

Machine Translation

美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件，威胁将会向机场等公众地方发动生化袭击後，关岛经保持高度戒备。

Machine translation is becoming very prevalent

Even PowerPoint has translation built into it!

The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport .

The American Guam international airport and the office will receive one to call self Saudi Arabian rich merchant Radden and so on the email which will send out, the threat can after public place launch biochemistry attacks and so on the airport, Guam after maintenance high alert.

2004: Which is the human?

Beijing Youth Daily said that under the Ministry of Agriculture, the beef will be destroyed after tests.

The Beijing Youth Daily pointed out that the seized beef would be disposed of after being examined according to advice from the Ministry of Agriculture.

?

2004: Which is the human?

Pakistan President Pervez Musharraf Wins Senate Confidence Vote

Pakistani President Musharraf Won the Trust Vote in Senate and Lower House

?

2004: Which is the human?

There was not a single vote against him."

No members vote against him. "

?

Warren Weaver (1947)

ingcmpnqsnwf cv fpn owoktvcv

hu ihgzsnwfv rqcffnw cw owgcnwf

kowazoanv ...

Warren Weaver (1947)

```
 e    e  e        e
ingcmpnqsnwf cv fpn owoktvcv
        e         e         e
hu ihgzsnwfv rqcffnw cw owgcnwf
       e
kowazoanv ...
```

Warren Weaver (1947)

```
 e    e  e      the
ingcmpnqsnwf cv fpn owoktvcv
        e         e         e
hu ihgzsnwfv rqcffnw cw owgcnwf
       e
kowazoanv ...
```

Warren Weaver (1947)

```
 e   he  e      the
ingcmpnqsnwf cv fpn owoktvcv
        e         e         e t
hu ihgzsnwfv rqcffnw cw owgcnwf
       e
kowazoanv ...
```

Warren Weaver (1947)

```
 e   he  e   of the
ingcmpnqsnwf cv fpn owoktvcv
        e         e         e t
hu ihgzsnwfv rqcffnw cw owgcnwf
       e
kowazoanv ...
```

Warren Weaver (1947)

e   he e   of the   fof
ingcmpnqsnwf cv fpn owoktvcv
e f   o e o   oe t
hu ihgzsnwfv rqcffnw cw owgcnwf
ef
kowazoanv ...

Warren Weaver (1947)

e   he e   ~~of~~ the
ingcmpnqsnwf cv fpn owoktvcv
e   e   e t
hu ihgzsnwfv rqcffnw cw owgcnwf
e
kowazoanv ...

Warren Weaver (1947)

e   he e   is the   sis
ingcmpnqsnwf cv fpn owoktvcv
e s   i e i   ie t
hu ihgzsnwfv rqcffnw cw owgcnwf
es
kowazoanv ...

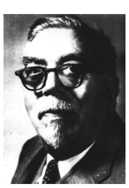
Warren Weaver (1947)

decipherment is the analysis
ingcmpnqsnwf cv fpn owoktvcv
of documents written in ancient
hu ihgzsnwfv rqcffnw cw owgcnwf
languages ...
kowazoanv ...

Warren Weaver (1947)

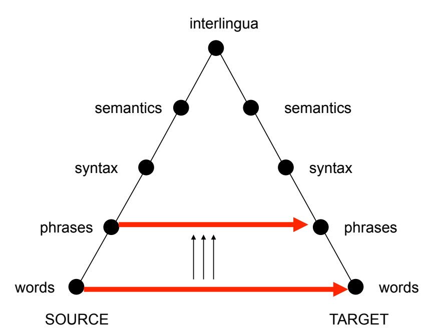
"When I look at an article in Russian, I say: this is really written in English, but it has been coded in some strange symbols. I will now proceed to decode."
- Warren Weaver, March 1947

"When I look at an article in Russian, I say: this is really written in English, but it has been coded in some strange symbols. I will now proceed to decode."
- Warren Weaver, March 1947

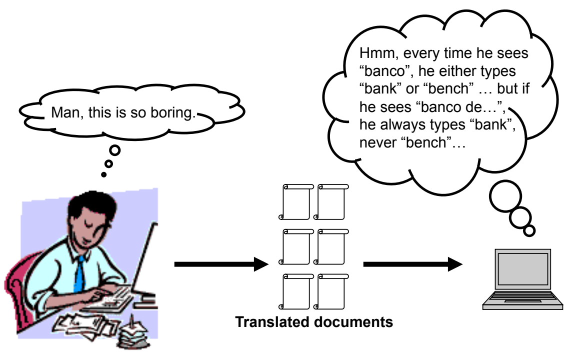


"... as to the problem of mechanical translation, I frankly am afraid that the [semantic] boundaries of words in different languages are too vague ... to make any quasi-mechanical translation scheme very hopeful."
- Norbert Wiener, April 1947

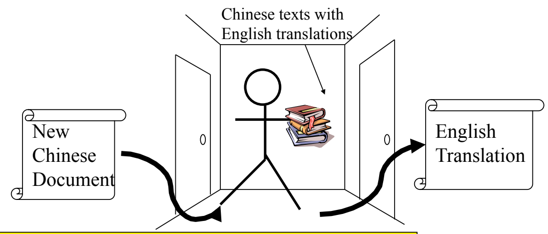
MT Pyramid



Data-Driven Machine Translation



Welcome to the Chinese Room



You can teach yourself to translate Chinese using *only* bilingual data (without grammar books, dictionaries, any people to answer your questions…)

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok **farok** ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok **farok** izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok **farok** ororok lalok sprok izok enemok . 7b. wat **jjat** bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok **farok** izok stok . 5b. totat **jjat** quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok **farok** ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok **farok** izok stok . 5b. totat jjat quat cat .	11a. lalok nok **crrrok** hihok yorok zanzanok . ??? 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok hihok yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok **hihok** ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok **hihok** yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok **hihok** mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok** yorok clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok **hihok** ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok **yorok** ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok **hihok** **yorok** zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok **hihok** mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok** clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok **clok** . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok** clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok **clok** . ??? 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok** clok kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok** **clok** kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . process of elimination 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat .
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok clok** kantok ok-yurp

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat . cognate?
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

Centauri/Arcturan [Knight, 1997]

Your assignment, put these words in order: **{ jjat, arrat, mat, bat, oloat, at-yurp }**

1a. ok-voon ororok sprok . 1b. at-voon bichat dat .	7a. lalok farok ororok lalok sprok izok enemok . 7b. wat jjat bichat wat dat vat eneat .
2a. ok-drubel ok-voon anok plok sprok . 2b. at-drubel at-voon pippat rrat dat .	8a. lalok brok anok plok nok . 8b. iat lat pippat rrat nnat .
3a. erok sprok izok hihok ghirok . 3b. totat dat arrat vat hilat .	9a. wiwok nok izok kantok ok-yurp . 9b. totat nnat quat oloat at-yurp .
4a. ok-voon anok drok brok jok . 4b. at-voon krat pippat sat lat .	10a. lalok mok nok yorok ghirok clok . 10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . 5b. totat jjat quat cat .	11a. lalok nok crrrok hihok yorok zanzanok . 11b. wat nnat arrat mat zanzanat . zero fertility
6a. lalok sprok izok jok stok . 6b. wat dat krat quat cat .	12a. lalok rarok nok izok hihok mok . 12b. wat nnat forat arrat vat gat .

It's Really Spanish/English

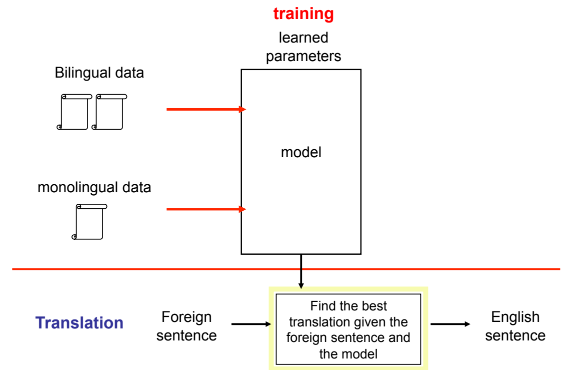
Clients do not sell pharmaceuticals in Europe => Clientes no venden medicinas en Europa

1a. Garcia and associates . 1b. Garcia y asociados .	7a. the clients and the associates are enemies . 7b. los clients y los asociados son enemigos .
2a. Carlos Garcia has three associates . 2b. Carlos Garcia tiene tres asociados .	8a. the company has three groups . 8b. la empresa tiene tres grupos .
3a. his associates are not strong . 3b. sus asociados no son fuertes .	9a. its groups are in Europe . 9b. sus grupos estan en Europa .
4a. Garcia has a company also . 4b. Garcia tambien tiene una empresa .	10a. the modern groups sell strong pharmaceuticals . 10b. los grupos modernos venden medicinas fuertes .
5a. its clients are angry . 5b. sus clientes estan enfadados .	11a. the groups do not sell zenzanine . 11b. los grupos no venden zanzanina .
6a. the associates are also angry . 6b. los asociados tambien estan enfadados .	12a. the small groups are not modern . 12b. los grupos pequenos no son modernos .

Data available

- Many languages
  - Europarl corpus has all European languages
    - http://www.statmt.org/europarl/
    - From a few hundred thousand sentences to a few million
  - French/English from French parliamentary proceedings
  - Lots of Chinese/English and Arabic/English from government projects/interests
    - Chinese-English: 440 million words (15-20 million sentence pairs)
    - Arabic-English: 790 million words (30-40 million sentence pairs)
  - Smaller corpora in many, many other languages
- Lots of monolingual data available in many languages
- Even less data with multiple translations available
- Available in limited domains
  - most data is either news or government proceedings
  - some other domains recently, like blogs

Statistical MT Overview

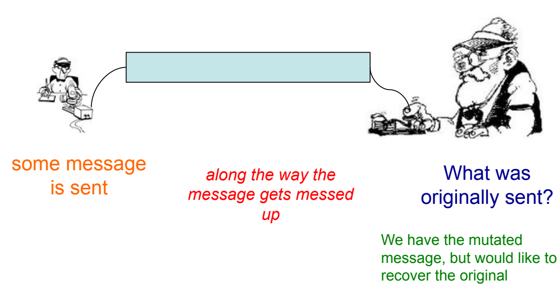


Statistical MT

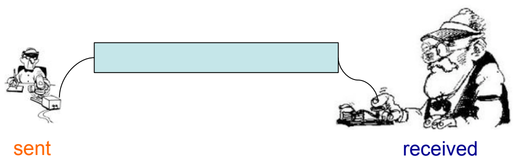
- We will model the translation process probabilistically
- Given a foreign sentence to translate, for any possible English sentence, we want to know what the probability that sentence is a translation of the foreign sentence
- If we can find the most probable English sentence, we're done

p(english sentence | foreign sentence)

Noisy channel model



Noisy channel model



model: p(sent | received)

Noisy channel model

Probabilistic model: $p(s \mid r)$

p(English| Foreign)

p(compressed| uncompressed)

p(simplified| unsimplified)

Given sentence pairs, gives us the probability

Noisy channel model

$$p(e \mid f) = \frac{p(f \mid e)\, p(e)}{p(f)}$$

Bayes' rule

$p(f)$ probability of the foreign sentence

$p(e)$ language model: what are likely English word sequences?

$p(f \mid e)$ translation model: how does the translation process happen? probability of the translated English sentence given the foreign sentence

Noisy channel model

$$p(e \mid f) = p(f \mid e)\, p(e)$$

Bayes' rule

~~$p(f)$ probability of the foreign sentence~~

$p(e)$ language model: what are likely English word sequences?

$p(f \mid e)$ translation model: how does the translation process happen? probability of the translated English sentence given the foreign sentence

Noisy channel model

model $p(e \mid f) \propto p(f \mid e)\, p(e)$

translation model — how do foreign sentences get translated to English sentences?

language model — what do English sentences look like?

Translation model

- The models define probabilities over inputs

$$p(f \mid e)$$

Morgen fliege ich nach Kanada zur Konferenz

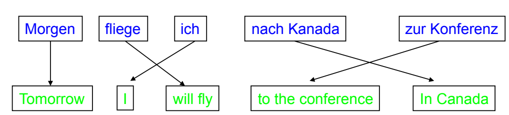
Tomorrow I will fly to the conference in Canada

What is the probability that the English sentence is a translation of the foreign sentence?

Translation model

- The models define probabilities over inputs

$$p(f \mid e)$$



- What is the probability of a foreign word being translated as a particular English word?
- What is the probability of a foreign foreign phrase being translated as a particular English phrase?
- What is the probability of a word/phrase changing ordering?
- What is the probability of a foreign word/phrase disappearing?
- What is the probability of a English word/phrase appearing?

Translation model

- The models define probabilities over inputs

$$p(f \mid e)$$

p(Morgen fliege ich nach Kanada zur Konferenz | Tomorrow I will fly to the conference in Canada) = 0.1

p(Morgen fliege ich nach Kanada zur Konferenz | I like peanut butter and jelly) = 0.0001

Language model

- The models define probabilities over inputs

$$p(e)$$

Tomorrow I will fly to the conference in Canada

What is a probability distribution?

- A probability distribution defines the probability over a space of possible inputs
- For the language model, what is the space of possible inputs?
  - A language model describes the probability over ALL possible combinations of English words
- For the translation model, what is the space of possible inputs?
  - ALL possible combinations of foreign words with ALL possible combinations of English words

One way to think about it…

Spanish (foreign) → Translation model → **Broken English** → language model → **English**

Que hambre tengo yo → What hunger have I, Hungry I am so, I am so hungry, Have I that hunger … → I am so hungry

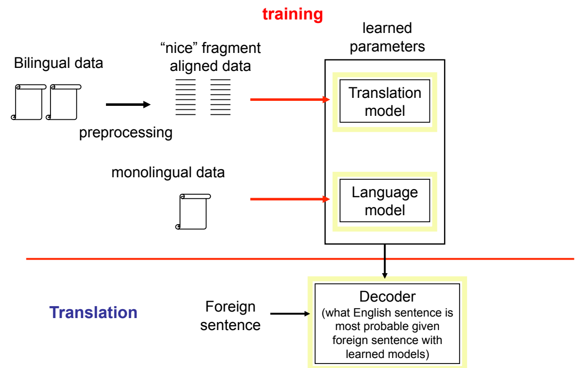
Translation

$$p(e \mid f) \propto p(f \mid e)\,p(e)$$

- Let's assume we have a translation model and a language model
- Given a foreign sentence, what question do we want to ask to translate that sentence into English?

$$\arg_e \max p(e \mid f) \propto p(f \mid e)\,p(e)$$

Statistical MT Overview



Basic Model, Revisited

$$\underset{e}{\text{argmax}}\; P(e \mid f) =$$

$$\underset{e}{\text{argmax}}\; P(e) \times P(f \mid e) / P(f) =$$

$$\underset{e}{\text{argmax}}\; P(e) \times P(f \mid e)$$

Basic Model, Revisited

$$\underset{e}{\text{argmax}}\; P(e \mid f) =$$

$$\underset{e}{\text{argmax}}\; P(e) \times P(f \mid e) / P(f) =$$

$$\underset{e}{\text{argmax}}\; P(e)^{2.4} \times P(f \mid e)$$ … works better!

Basic Model, Revisited

$$\underset{e}{\text{argmax}}\ P(e \mid f) =$$

$$\underset{e}{\text{argmax}}\ P(e) \times P(f \mid e) / P(f)$$

$$\underset{e}{\text{argmax}}\ P(e)^{2.4} \times P(f \mid e) \times \text{length}(e)^{1.1}$$

Rewards longer hypotheses, since these are unfairly punished by P(e)

Basic Model, Revisited

$$\underset{e}{\text{argmax}}\ \underbrace{P(e)^{2.4} \times P(f \mid e) \times \text{length}(e)^{1.1} \times KS^{3.7} \ldots}$$

Lots of knowledge sources vote on any given hypothesis.

"Knowledge source" = "feature function" = "score component".

A feature function simply scores a hypothesis with a real value.

(May be binary, as in "e has a verb").

Problems for Statistical MT

- Preprocessing
  - How do we get aligned bilingual text?
  - Tokenization
  - Segmentation (document, sentence, word)
- Language modeling
  - Given an English string e, assigns P(e) by formula
- Translation modeling
  - Given a pair of strings <f,e>, assigns P(f | e) by formula
- Decoding
  - Given a language model, a translation model, and a new sentence f … find translation e maximizing P(e) * P(f | e)
- Parameter optimization
  - Given a model with multiple feature functions, how are they related? What are the optimal paraeters?
- Evaluation
  - How well is a system doing? How can we compare two systems?

Problems for Statistical MT

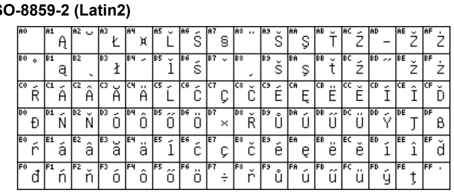
- Preprocessing
- Language modeling
- Translation modeling
- Decoding
- Parameter optimization
- Evaluation

From No Data to Sentence Pairs

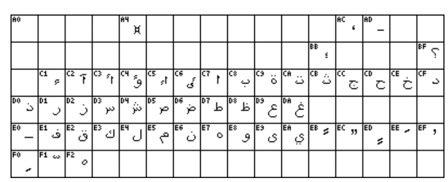
- Easy way: Linguistic Data Consortium (LDC)
- Really hard way: pay $$$
  - Suppose one billion words of parallel data were sufficient
  - At 20 cents/word, that's $200 million
- Pretty hard way: Find it, and then earn it!
  - De-formatting
  - Remove strange characters
  - Character code conversion
  - **Document alignment**
  - **Sentence alignment**
  - **Tokenization (also called Segmentation)**

If you don't get the characters right…

ISO-8859-2 (Latin2)



ISO-8859-6 (Arabic)

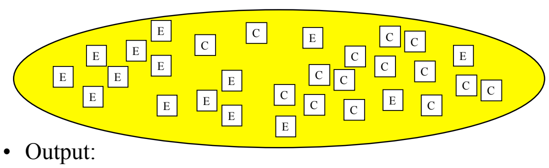


Chinese?

- **GB Code**
- **GBK Code**
- **Big 5 Code**
- **CNS-11643-1992**
- …

Document Alignment

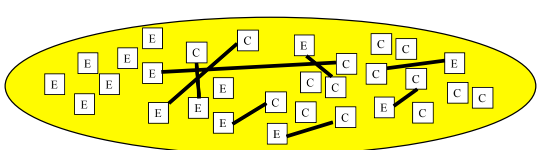
- Input:
  - Big bag of files obtained from somewhere, believed to contain pairs of files that are translations of each other.



- Output:
  - List of pairs of files that are actually translations.

Document Alignment

- Input:
  - Big bag of files obtained from somewhere, believed to contain pairs of files that are translations of each other.



- Output:
  - List of pairs of files that are actually translations.

Sentence Alignment

The old man is happy. He has fished many times. His wife talks to him. The fish are jumping. The sharks await.	El viejo está feliz porque ha pescado muchos veces. Su mujer habla con él. Los tiburones esperan.

Sentence Alignment

1. The old man is happy.
2. He has fished many times.
3. His wife talks to him.
4. The fish are jumping.
5. The sharks await.

1. El viejo está feliz porque ha pescado muchos veces.
2. Su mujer habla con él.
3. Los tiburones esperan.

Sentence Alignment

1. The old man is happy.
2. He has fished many times.
3. His wife talks to him.
4. The fish are jumping.
5. The sharks await.

1. El viejo está feliz porque ha pescado muchos veces.
2. Su mujer habla con él.
3. Los tiburones esperan.

Sentence Alignment

1. The old man is happy. He has fished many times. — 1. El viejo está feliz porque ha pescado muchos veces.
2. His wife talks to him. — 2. Su mujer habla con él.
3. The sharks await. — 3. Los tiburones esperan.

Note that unaligned sentences are thrown out, and sentences are merged in n-to-m alignments (n, m > 0).

Tokenization (or Segmentation)

- English
  - Input (some byte stream):

    `"There," said Bob.`
  - Output (7 “tokens” or “words”):

    `" There , " said Bob .`
- Chinese
  - Input (byte stream): 美国关岛国际机场及其办公室均接获一名自称沙地阿拉伯富商拉登等发出的电子邮件。
  - Output: 美国 关岛国 际机 场 及其 办公 室 均接获 一名 自称 沙地 阿拉 伯富 商拉登 等发 出 的 电子邮件。

Problems for Statistical MT

- Preprocessing
- Language modeling
- Translation modeling
- Decoding
- Parameter optimization
- Evaluation

Language Modeling

- Most common: n-gram language models
- More data the better (Google n-grams)
- Domain is important

Problems for Statistical MT

- Preprocessing
- Language modeling
- Translation modeling
- Decoding
- Parameter optimization
- Evaluation

Translation Model

Learn How to Translate from Data

Direct Estimation:

Mary did not slap the green witch

not enough data for this
(most input sentences unseen)

Maria no dió una botefada a la bruja verde

Generative Model

Break up process into smaller steps:

Mary did not slap the green witch

sufficient statistics for smaller steps

Maria no dió una botefada a la bruja verde

What kind of Translation Model?

May use syntactic and semantic representations:

Mary did not slap the green witch

Word-level models

Phrasal models

Syntactic models

Semantic models

Maria no dió una botefada a la bruja verde

The Classic Translation Model

Word Substitution/Permutation [IBM Model 3, Brown et al., 1993]

Generative story:

Mary did not slap the green witch

Mary not slap slap slap the green witch — n(3|slap)

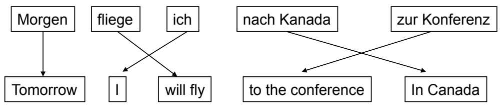
Mary not slap slap slap NULL the green witch — P-Null

Maria no dió una botefada a la verde bruja — t(la|the)

Maria no dió una botefada a la bruja verde — d(j|i)

Probabilities can be learned from raw bilingual text.

Phrase-Based Statistical MT

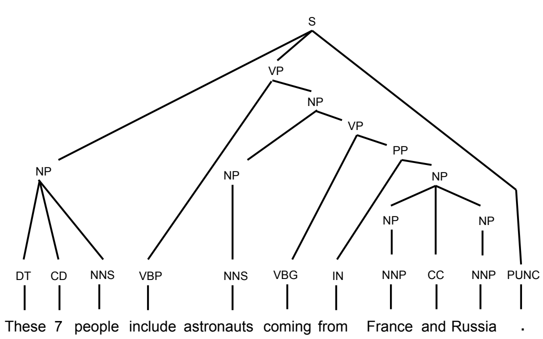


- Foreign input segmented in to phrases
  - "phrase" is any sequence of words
- Each phrase is probabilistically translated into English
  - P(to the conference | zur Konferenz)
  - P(into the meeting | zur Konferenz)
- Phrases are probabilistically re-ordered
- See [Koehn et al, 2003] for an intro.

Advantages of Phrase-Based

- Many-to-many mappings can handle non-compositional phrases
- Easy to understand
- Local context is very useful for disambiguating
  - "Interest rate" → …
  - "Interest in" → …
- The more data, the longer the learned phrases
  - Sometimes whole sentences

Syntax



Problems for Statistical MT

- Preprocessing
- Language modeling
- Translation modeling
- Decoding
- Parameter optimization
- Evaluation

Decoding

- Of all conceivable English word strings, find the one maximizing P(e) x P(f | e)
- Decoding is an NP-complete problem (for many translation models
  - (Knight, 1999)
- Several decoding strategies are often available

Search

$\arg_e \max p(f \mid e)\, p(e)$

$f_1\ f_2\ f_3\ f_4\ f_5\ f_6$

Search

$\arg_e \max p(f \mid e)\, p(e)$

$f_1\ f_2\ f_3\ f_4\ f_5\ f_6$ →

partial trans.
$f_1\ f_2\ f_3\ f_4\ f_5\ f_6$

partial trans.
$\mathbf{f_1\ f_2\ f_3\ f_4\ f_5\ f_6}$

…

partial trans.
$f_1\ f_2\ f_3\ f_4\ f_5\ f_6$

Search

$\arg_e \max p(f \mid e)p(e)$

$f_1\ f_2\ f_3\ f_4\ f_5\ f_6$

partial trans.
$f_1\ f_2\ f_3\ f_4\ f_5\ f_6$

partial trans.
$f_1\ f_2\ f_3\ f_4\ f_5\ f_6$

partial trans.
$f_1\ f_2\ f_3\ f_4\ f_5\ f_6$

partial trans.
$f_1\ f_2\ f_3\ f_4\ f_5\ f_6$

partial trans.
$\mathbf{f_1\ f_2\ f_3\ f_4\ f_5\ f_6}$

…

…

partial trans.
$f_1\ f_2\ f_3\ f_4\ f_5\ f_6$

Search

$\arg_e \max p(f \mid e)p(e)$

$f_1\ f_2\ f_3\ f_4\ f_5\ f_6$

partial trans.
$f_1\ f_2\ f_3\ f_4\ f_5\ f_6$

partial trans.
$f_1\ f_2\ f_3\ f_4\ f_5\ f_6$

partial trans.
$f_1\ f_2\ f_3\ f_4\ f_5\ f_6$

partial trans.
$f_1\ f_2\ f_3\ f_4\ f_5\ f_6$

partial trans.
$\mathbf{f_1\ f_2\ f_3\ f_4\ f_5\ f_6}$

…

…

partial trans.
$f_1\ f_2\ f_3\ f_4\ f_5\ f_6$

- State space search problem
- Huge space
- Evaluate intermediate hypotheses using model and prune partial hypotheses

Problems for Statistical MT

- Preprocessing
- Language modeling
- Translation modeling
- Decoding
- Parameter optimization
- Evaluation

The Problem: Learn Lambdas

$$p(e \mid f) = \frac{p(f \mid e)p(e)}{p(f)}$$

$$= \frac{p(f \mid e)^{\lambda_1} p(e)^{\lambda_2}}{\sum_{e'} p(f \mid e')^{\lambda_1} \lambda_2 p(e')^{\lambda_2}}$$

$$= \frac{p(f \mid e)^{\lambda_1} p(e)^{\lambda_2} p(e \mid f)^{\lambda_3} length(e)^{\lambda_4} \ldots}{\sum_{e'} p(f \mid e')^{\lambda_1} p(e')^{\lambda_2} p(e' \mid f)^{\lambda_3} length(e')^{\lambda_4} \ldots}$$

$$= \frac{\exp(\lambda_1 \log p(f \mid e) + \lambda_2 \log p(e) + \lambda_3 \log p(e \mid f) + \lambda_4 length(e) \ldots)}{\sum_{e'} \exp(\lambda_1 \log p(f \mid e') + \lambda_2 \log p(e') + \lambda_3 \log p(e' \mid f) + \lambda_4 length(e') \ldots)}$$

$$= \frac{\exp\left(\sum_i \lambda_i h_i(f, e)\right)}{\sum_{e'} \exp\left(\sum_i \lambda_i h_i(f, e')\right)}$$

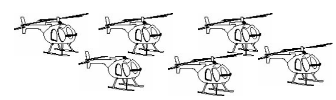
Given a data set with foreign/English sentences, find the λ's that:
- maximize the likelihood of the data
- maximize an evaluation criterion

Problems for Statistical MT

- Preprocessing
- Language modeling
- Translation modeling
- Decoding
- Parameter optimization
- Evaluation

MT Evaluation

- Source only
- Manual:
  - SSER (subjective sentence error rate)
  - Correct/Incorrect
  - Error categorization
- Objective usage testing



- Automatic:
  - WER (word error rate)
  - BLEU (Bilingual Evaluation Understudy)
  - NIST
  - Named-Entity

Reference Evaluation

Reference (human) translation:
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/ chemical attack against public places such as the airport .

Machine translation:
The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

BLEU Evaluation Metric

(Papineni et al, ACL-2002)

Reference (human) translation:
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/ chemical attack against public places such as the airport .

Machine translation:
The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

- N-gram precision (score is between 0 & 1)
  - What percentage of machine n-grams can be found in the reference translation?
    - An n-gram is an sequence of n words
  - Not allowed to use same portion of reference translation twice (can't cheat by typing out "the the the the the")
- Brevity penalty
  - Can't just type out single word "the" (precision 1.0!)

*** Amazingly hard to "game" the system (i.e., find a way to change machine output so that BLEU goes up, but quality doesn't)

BLEU Evaluation Metric

(Papineni et al, ACL-2002)

Reference (human) translation:
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/ chemical attack against public places such as the airport .

Machine translation:
The American [?] international airport and its the office all receives one calls self the sand Arab rich business [?] and so on electronic mail , which sends out ; The threat will be able after public place and so on the airport to start the biochemistry attack , [?] highly alerts after the maintenance.

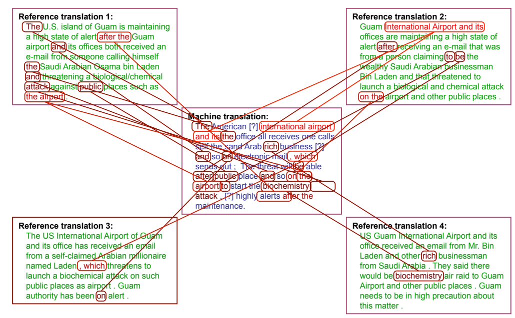
- BLEU formula
  - Generally N=4
  - w_i=1/N (uniform weights)

$$BLEU = \prod_{i=1}^{N} \sqrt[N]{p_i^{w_i}}\, BP$$

BP=brevity penalty
p_i=i-gram precision

$$BLEU = BP \cdot \exp\left(\sum_{i=1}^{N} w_i \log p_i \right)$$

Multiple Reference Translations



Available Resources

- Bilingual corpora
  - 100m+ words of Chinese/English and Arabic/English, LDC (www.ldc.upenn.edu)
  - Lots of French/English, Spanish/French/English, LDC
  - European Parliament (sentence-aligned), 11 languages, Philipp Koehn, ISI
    - (www.isi.edu/~koehn/publications/europarl)
  - 20m words (sentence-aligned) of English/French, Ulrich Germann, ISI
    - (www.isi.edu/natural-language/download/hansard/)
- Sentence alignment
  - Dan Melamed, NYU (www.cs.nyu.edu/~melamed/GMA/docs/README.htm)
  - Xiaoyi Ma, LDC (Champollion)
- Word alignment
  - GIZA, JHU Workshop '99 (www.clsp.jhu.edu/ws99/projects/mt/)
  - GIZA++, RWTH Aachen (www-i6.Informatik.RWTH-Aachen.de/web/Software/GIZA++.html)
  - Manually word-aligned test corpus (500 French/English sentence pairs), RWTH Aachen
  - Shared task, NAACL-HLT'03 workshop
- Decoding
  - ISI ReWrite Model 4 decoder (www.isi.edu/licensed-sw/rewrite-decoder/)
  - ISI Pharaoh phrase-based decoder
- Statistical MT Tutorial Workbook, ISI (www.isi.edu/~knight/)
- Annual common-data evaluation, NIST (www.nist.gov/speech/tests/mt/index.htm)

Some Papers Referenced on Slides

- ACL
  - [Och, Tillmann, & Ney, 1999]
  - [Och & Ney, 2000]
  - [Germann et al, 2001]
  - [Yamada & Knight, 2001, 2002]
  - [Papineni et al, 2002]
  - [Alshawi et al, 1998]
  - [Collins, 1997]
  - [Koehn & Knight, 2003]
  - [Al-Onaizan & Knight, 2002]
  - [Och & Ney, 2002]
  - [Och, 2003]
  - [Koehn et al, 2003]
- EMNLP
  - [Marcu & Wong, 2002]
  - [Fox, 2002]
  - [Munteanu & Marcu, 2002]
- AI Magazine
  - [Knight, 1997]
- www.isi.edu/~knight
  - [MT Tutorial Workbook]
- AMTA
  - [Soricut et al, 2002]
  - [Al-Onaizan & Knight, 1998]
- EACL
  - [Cmejrek et al, 2003]
- Computational Linguistics
  - [Brown et al, 1993]
  - [Knight, 1999]
  - [Wu, 1997]
- AAAI
  - [Koehn & Knight, 2000]
- IWNLG
  - [Habash, 2002]
- MT Summit
  - [Charniak, Knight, Yamada, 2003]
- NAACL
  - [Koehn, Marcu, Och, 2003]
  - [Germann, 2003]
  - [Graehl & Knight, 2004]
  - [Galley, Hopkins, Knight, Marcu, 2004]