

Introduction to Statistical Machine Translation

David Kauchak
CS159 – Fall 2020

Philipp Koehn
School of Informatics
University of Edinburgh

Some slides adapted from
Kevin Knight
USC/Information Sciences Institute
USC/Computer Science Department

Dan Klein
Computer Science Department
UC Berkeley

1

Admin

Assignment 5 out

Quiz #2

2

Language translation



3

MT Systems

Where have you seen machine translation systems?

4

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裔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。



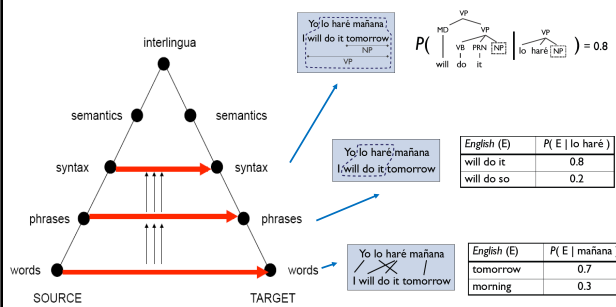
The U.S. island of Guam is maintaining a high state of alert after the Guam airport and its offices both received an e-mail from someone calling himself the Saudi Arabian Osama bin Laden and threatening a biological/chemical attack against public places such as the airport.

A good test for natural language processing.

Requires capabilities in both interpretation and generation.

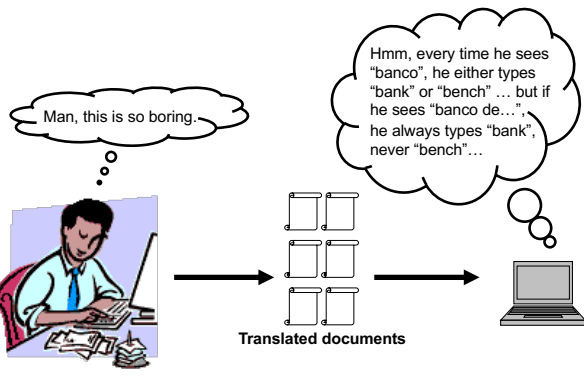
5

Levels of Transfer



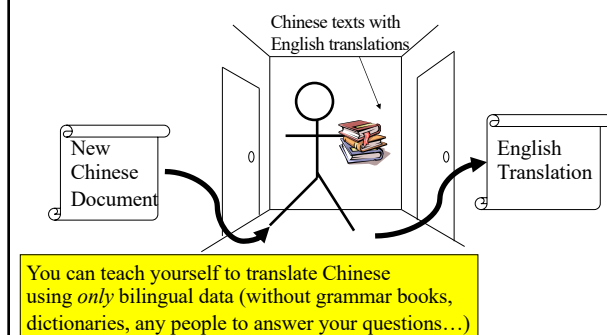
6

Data-Driven Machine Translation



9

Welcome to the Chinese Room



10

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok errok hihok yorok klok kantok ok-yurp

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

11

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok errok hihok yorok klok kantok ok-yurp

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

12

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok errok hihok yorok klok kantok ok-yurp

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

13

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok errok hihok yorok klok kantok ok-yurp

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

14

Centauri/Arcturan [Knight, 1997]

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

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

15

Centauri/Arcturan [Knight, 1997]

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

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

16

Centauri/Arcturan [Knight, 1997]

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

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

17

Centauri/Arcturan [Knight, 1997]

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

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

18

Centauri/Arcturan [Knight, 1997]

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

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

19

Centauri/Arcturan [Knight, 1997]

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

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

20

Centauri/Arcturan [Knight, 1997]

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

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

21

Centauri/Arcturan [Knight, 1997]

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

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

22

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 clientes 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 .

23

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: Hundreds of millions of sentence pairs)
 - Arabic-English: ~One hundred million sentence pairs
- Smaller corpora in many, many other languages

Lots of monolingual data available in many languages

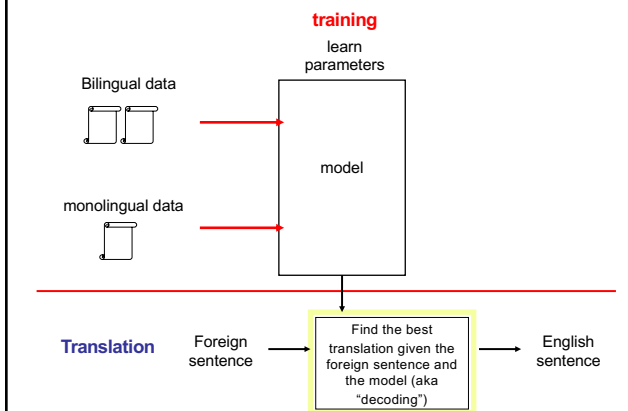
Even less data with multiple translations available

Available in limited domains

- most data is either news or government proceedings
- some other domains recently, like blogs

24

Statistical MT Overview



25

Statistical MT

We will model the translation process probabilistically

Given a foreign sentence to translate, for any possible English sentence, we want to know the probability that the sentence is a translation of the foreign sentence

If we can find the most probable English sentence, we're done

$$p(\text{english sentence} \mid \text{foreign sentence})$$

26

Translation

Probabilistic model: $p(e|f)$ $p(\text{English} | \text{Foreign})$

What is the translation problem then?

$$\text{translation}(f) = \arg_e \max p(e|f)$$

27

Noisy channel model

$$p(e|f) = \frac{p(f|e)p(e)}{p(f)} \quad \text{Bayes' rule}$$

$p(f)$ probability of the foreign sentence

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

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

28

Noisy channel model

$$p(e|f) = p(f|e)p(e) \quad \text{Bayes' rule}$$

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

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

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

29

Noisy channel model

$$p(e|f) = p(f|e)p(e) \quad \text{Bayes' rule}$$

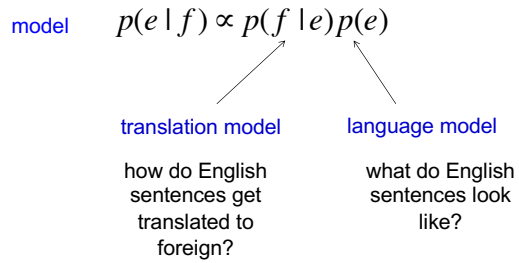
~~$p(f)$ probability of the foreign sentence why?~~

$$\text{translation}(f) = \arg_e \max \frac{p(f|e)p(e)}{p(f)} = \arg_e \max p(f|e)p(e)$$

this is a constant for any given f

30

Noisy channel model



31

Translation model

The models define probabilities over inputs

$$p(f | e)$$

Morgen fliege ich nach Kanada zur Konferenz

Tomorrow I will fly to the conference in Canada

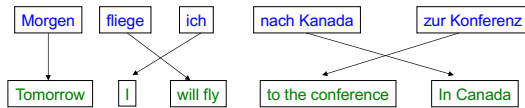
What is the probability that the English sentence is a translation of the foreign sentence?

32

Translation model

The models define probabilities over inputs

$$p(f | e)$$



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

33

Translation model

The models define probabilities over inputs

$$p(f | e)$$

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

$$p(\text{Morgen fliege ich nach Kanada zur Konferenz} | \text{I like peanut butter and jelly}) = 0.0001$$

34

Language model

The models define probabilities over inputs

$$p(e)$$

Tomorrow I will fly to the conference in Canada

35

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?

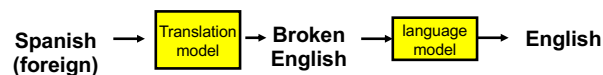
- A language model describes the probability over **ALL** possible combinations of English words

For the translation model, what is the space of possible inputs?

- **ALL** possible combinations of foreign words with **ALL** possible combinations of English words

36

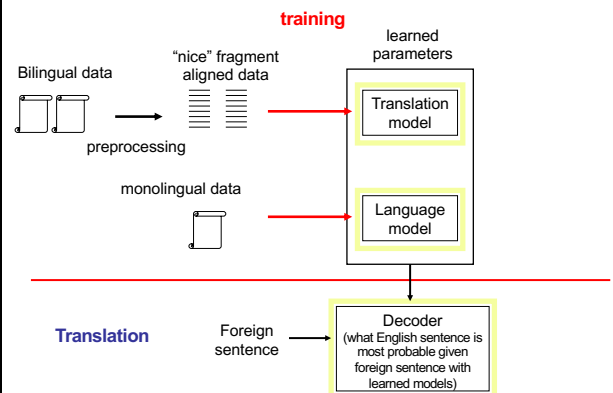
One way to think about it...



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

37

Statistical MT Overview



38

Basic Model, Revisited

$$\operatorname{argmax}_e P(e | f) =$$

$$\operatorname{argmax}_e P(e) \times P(f | e) / P(f) =$$

$$\operatorname{argmax}_e P(e) \times P(f | e)$$

39

Basic Model, Revisited

$$\operatorname{argmax}_e P(e | f) =$$

$$\operatorname{argmax}_e P(e) \times P(f | e) / P(f) =$$

$$\operatorname{argmax}_e P(e)^{2.4} \times P(f | e) \quad \dots \text{ works better!}$$

40

Basic Model, Revisited

$$\operatorname{argmax}_e P(e | f) =$$

$$\operatorname{argmax}_e P(e) \times P(f | e) / P(f)$$

$$\operatorname{argmax}_e P(e)^{2.4} \times P(f | e) \times \text{length}(e)^{1.1}$$

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

41

Basic Model, Revisited

$$\operatorname{argmax}_e P(e)^{2.4} \times P(f | e) \times \text{length}(e)^{1.1} \times \text{KS}^{3.7} \dots$$

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").

42

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 $\langle f, e \rangle$, 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 parameters?

Evaluation

- How well is a system doing? How can we compare two systems?

43