

CS-AD 220 – Spring 2016

Natural Language Processing

Session 21: 14-Apr-16

Prof. Nizar Habash

Slides from Koehn, Knight, Hoang, and others with extensions

NYUAD Course CS-AD 220 – Spring 2016

Natural Language Processing

Assignment #3 : POS Tagging and Parsing

Assigned Mar 31, 2016 / Due Apr 17, 2016 (11:59pm)

I. Grading & Submission

This assignment is about the development of a dependency parser and a part-of-speech (POS) tagger for English. The assignment accounts for 15% of the full grade. It consists of five exercises. **There is also a bonus exercise that can count for up to 5% of the full grade.** The additional exercise consists of a parsing competition on an unseen test set. Participation earns 2%. The first, second and third ranked systems earn additional 3%, 2% and 1%, respectively.

Assignment #3 posted on NYU Classes

START EARLY!

DEADLINE PUSHED FORWARD TO APR 17

Looking ahead

- Hackathon!
 - April 15-17
- Deadline Assignment #3
 - April 17
- Prof. Jan Hajic
 - April 21

Looking ahead

- April 19
 - Assignment #4 to be given out
 - Come prepared with your MT team
 - 2-3 people
 - Have a cool name!
 - Bring a laptop (1 per team at least)
 - We will do a hands on session with Dr. Hoang

Probability Refresher

- Calculating probability

$$p(A = x) = \frac{\text{count}(A = x)}{\sum_i \text{count}(A = i)}$$

- Joint Probability

$$p(A = x, B = y) = p(A = x)p(B = y)$$

- Marginalization

$$p(A = x) = \sum_i p(A = x, B = i)$$

- Conditional Probability

- Chain Rule

$$p(A = x, B = y) = p(A = x | B = y)p(B = y)$$

- Bayes Rule

$$p(A = x | B = y) = \frac{p(A = x, B = y)}{p(B = y)}$$

- Calculating conditional probability

$$p(A = x | B = y) = \frac{\text{count}(A = x, B = y)}{\sum_i \text{count}(A = i, B = y)}$$

Probability Refresher

- Calculating probability

$$p(x) = \frac{\text{count}(x)}{\sum_i \text{count}(i)}$$

- Joint Probability

$$p(x, y) = p(x)p(y)$$

- Marginalization

$$p(x) = \sum_i p(x, i)$$

- Conditional Probability

- Chain Rule

$$p(x, y) = p(x | y)p(y)$$

- Bayes Rule

$$p(x | y) = \frac{p(y | x)p(x)}{p(y)}$$

- Calculating conditional probability

$$p(x | y) = \frac{\text{count}(x, y)}{\sum_i \text{count}(i, y)}$$

Word Alignment

The Big Picture

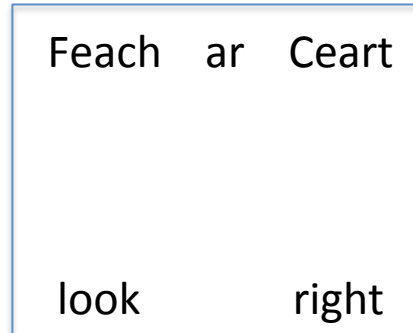
- What we have
 - Parallel text



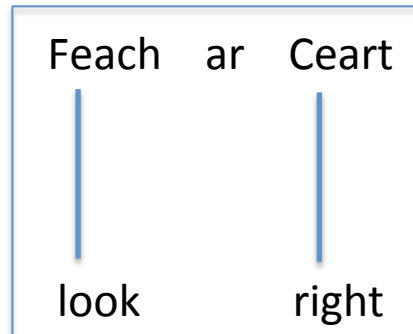
- What we want
 - Translation rules
 - Probabilities
 - Feach → look ,with probability ???
 - Feach ar → look ,with probability ???
 - Ceart → right ,with probability ???
 -

Word Alignment

- Parallel text



- With alignment



Centauri-Arcturan Parallel Text

1a. ok-voon ororok sprok .

1b. at-voon bichat dat .

2a. ok-drubel ok-voon anak plok sprok .

2b. at-drubel at-voon pippat rrat dat .

3a. erok sprok izok hihok ghrok .

3b. totat dat arrat vat hilat .

4a. ok-voon anak drok brok jok .

4b. at-voon krat pippat sat lat .

5a. wiwok farok izok stok .

5b. totat jjat quat cat .

6a. lalok sprok izok jok stok .

6b. wat dat krat quat cat .

7a. lalok farok ororok lalok sprok izok enemok .

7b. wat jjat bichat wat dat vat eneat .

8a. lalok brok anak plok nok .

8b. iat lat pippat rrat nnat .

9a. wiwok nok izok kantok ok-yurp .

9b. totat nnat quat oloat at-yurp .

10a. lalok mok nok yorok ghrok klok .

10b. wat nnat gat mat bat hilat .

11a. lalok nok crrrok hihok yorok zanzanok .

11b. wat nnat arrat mat zanzanat .

12a. lalok rarok nok izok hihok mok .

12b. wat nnat forat arrat vat gat .

(from Knight (1997): Automating Knowledge Acquisition for Machine Translation)

Centauri/Arcturan [Knight, 1997]

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

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrrok 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 ghrok .	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 ghrok klok .
4b. at-voon krat pippat sat lat .	10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok .	11a. lalok nok crrrok 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 .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: farok crrok 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 ghrok .	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 ghrok klok .
4b. at-voon krat pippat sat lat .	10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok .	11a. lalok nok crrok 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 .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrok 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 eneant .
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 ghrok .	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 ghrok klok .
4b. at-voon krat pippat sat lat .	10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok .	11a. lalok nok crrok 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 .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** **crrok** 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 ghrok .	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 ghrok klok .
4b. at-voon krat pippat sat lat .	10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok .	11a. lalok nok crrok 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 .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **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 ghrok .	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 ghrok klok .
4b. at-voon krat pippat sat lat .	10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok .	11a. lalok nok crrrok hihok yorok zanzanok .
/	11b. wat nnat arrat mat zanzanat .
5b. totat jjat quat cat .	12a. lalok rarok nok izok hihok mok .
6a. lalok sprok izok jok stok .	12b. wat nnat forat arrat vat gat .
6b. wat dat krat quat cat .	

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **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 ghrok .	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 ghrok klok .
4b. at-voon krat pippat sat lat .	10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok .	11a. lalok nok crrrok 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 .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok** **clock** 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 ghrok .	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 ghrok clock .
4b. at-voon krat pippat sat lat .	10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok .	11a. lalok nok crrrok 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 .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok** **clock** 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 clock .
4b. at-voon krat pippat sat lat .	10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok .	11a. lalok nok crrrok 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 .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok yorok** **clock** 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 ghrok . / /	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 ghrok clock . / / /
4b. at-voon krat pippat sat lat .	10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok . /	11a. lalok nok crrrok 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 .

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** crrrok **hihok** **yorok** **clock** 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 ghrok .	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 ghrok clock .
4b. at-voon krat pippat sat lat .	10b. wat nnat gat mat bat hilat .
5a. wiwok farok izok stok .	11a. lalok nok crrrok 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 .

process of
elimination

Centauri/Arcturan [Knight, 1997]

Your assignment, translate this to Arcturan: **farok** **crrrok** **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 ghrok . / /	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 ghrok klok . / /
4b. at-voon krat pippat sat lat .	10b. wat nnat gat mat bat hilat . / /
5a. wiwok farok izok stok . /	11a. lalok nok crrrok 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 . / /

cognate?

Centauri/Arcturan [Knight, 1997]

Your assignment, put these words in order: { jjat, arrat, mat, bat, oloat, 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 ghirook . / /	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 ghirook klok . / / /
4b. at-voon krat pippat sat lat .	10b. wat nnat gat mat bat hilat . / / /
5a. wiwok farok izok stok . /	11a. lalok nok crrrok hihok yorok zanzanok . / zero
5b. totat jjat quat cat .	11b. wat nnat arrat mat zanzanat . fertility
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 .

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 .

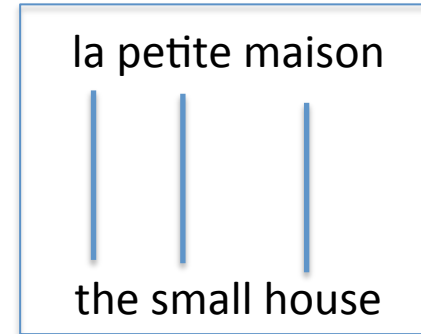
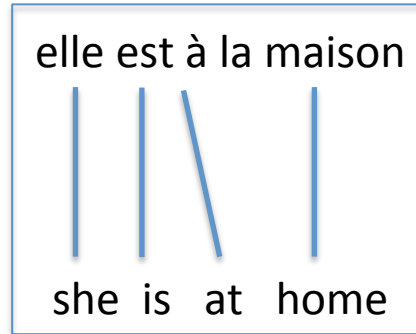
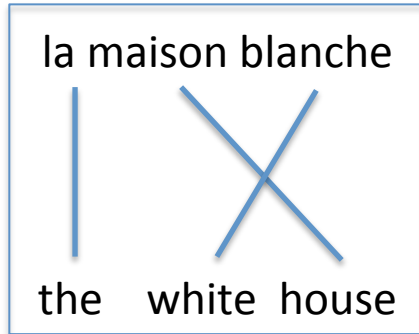
11a. the groups do not sell zenzanine .
11b. los grupos no venden zanzanina .

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 .

Compute Translation Rules

Parallel Corpus



Counts

Count(la, the) = 2

Count(maison, house) = 2

Count(maison, home) = 1

Count(blanche, white) = 1

Count(elle, she) = 1

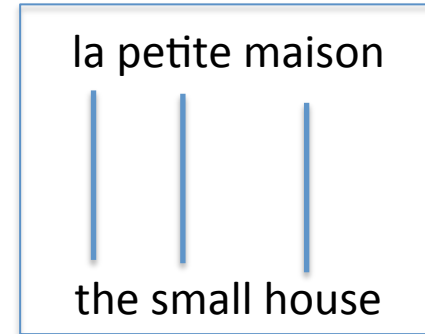
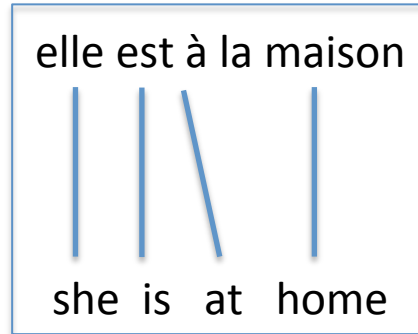
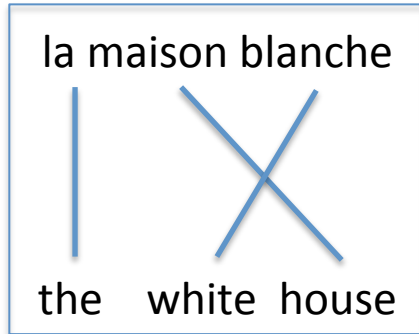
Count(est, is) = 1

Count(à, at) = 1

Count(petite, small) = 1

Compute Translation Rules

Parallel Corpus



Counts

Count(la, the) = 2

Count(maison, house) = 2

Count(maison, home) = 1

Count(blanche, white) = 1

Count(elle, she) = 1

Count(est, is) = 1

Count(à, at) = 1

Count(petite, small) = 1

Translation Rules

la → the , probability 1.0

maison → house , probability 0.66666

maison → home , probability 0.33333

blanche → white , probability 1.0

elle → she , probability 1.0

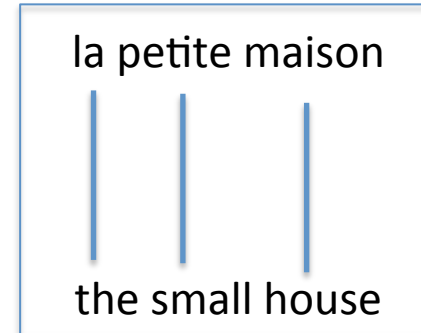
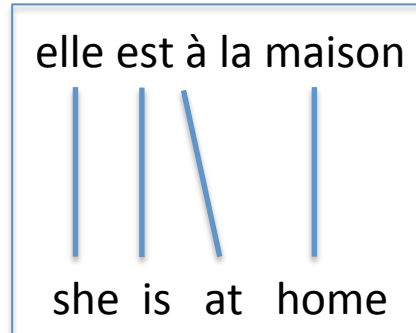
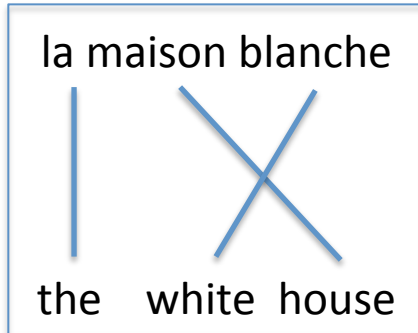
est → is , probability 1.0

à → at , probability 1.0

petite → small , probability 1.0

Word Alignment

Parallel Corpus



Counts

Count(la, the) = 2

Count(maison, house) = 2

Count(maison, home) = 1

Count(blanche, white) = 1

Count(elle, she) = 1

Count(est, is) = 1

Count(à, at) = 1

Count(petite, small) = 1

Translation Rules

$t(\text{the} | \text{la}) = 1.0$

$t(\text{house} | \text{maison}) = 0.66666$

$t(\text{home} | \text{maison}) = 0.33333$

$t(\text{white} | \text{blanche}) = 1.0$

$t(\text{she} | \text{elle}) = 1.0$

$t(\text{is} | \text{est}) = 1.0$

$t(\text{at} | \text{à}) = 1.0$

$t(\text{small} | \text{petite}) = 1.0$

Word Alignment

- Given parallel corpus
 - With word alignment
 - Compute translation rules
- Given translation rules
 - Compute probability of translation

Word-Based Translation Model

Translation Rules

$t(\text{the}|\text{la}) = 1.0$

$t(\text{house}|\text{maison}) = 0.66666$

$t(\text{home}|\text{maison}) = 0.33333$

$t(\text{is}|\text{est}) = 1$

$t(\text{small}|\text{petite}) = 1$

Calculating Probability of a Translation + Implied Alignment

la maison blanche est petite

the white house is small

Word-Based Translation Model

Translation Rules

$$t(\text{the}|\text{la}) = 1.0$$

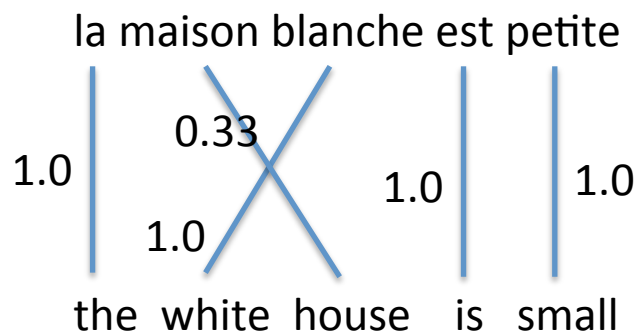
$$t(\text{house}|\text{maison}) = 0.66666$$

$$t(\text{home}|\text{maison}) = 0.33333$$

$$t(\text{is}|\text{est}) = 1$$

$$t(\text{small}|\text{petite}) = 1$$

Calculating Probability of a Translation + Implied Alignment



Word-Based Translation Model

Translation Rules

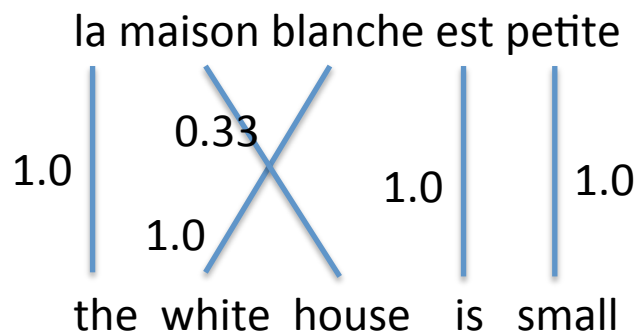
$$t(\text{the}|\text{la}) = 1.0$$

$$t(\text{house}|\text{maison}) = 0.66666$$

$$t(\text{home}|\text{maison}) = 0.33333$$

.....

Calculating Probability of a Translation + Implied Alignment



$$\begin{aligned} p(\text{the white house is small, alignment} \mid \text{la maison ...}) \\ &= p(\text{the}|\text{la}) * p(\text{white}|\text{blanche}) * p(\text{house}|\text{maison})..... \\ &= 1.0 \times 1.0 \times 0.33 \times 1.0 \times 1.0 \\ &= 0.33 \end{aligned}$$

Word-Based Translation Model

Translation Rules

$t(\text{the}|\text{la}) = 1.0$

$t(\text{house}|\text{maison}) = 0.66666$

$t(\text{home}|\text{maison}) = 0.33333$

.....

$t(\text{white}|\text{maison}) = 0.1$

$t(\text{house}|\text{blanche}) = 0.2$

Calculating Probability of a Translation + Implied Alignment



$$\begin{aligned} p(\text{the white house is small, alignments} | \text{la maison....}) \\ &= 1.0 \times 0.1 \times 0.2 \times 1 \times 1 \\ &= 0.02 \end{aligned}$$

Definitions

Alignment Function

- Formalizing alignment with an alignment **function**
 - Map
 - source word at position i
 - Target word at position j
- $a: i \rightarrow j$
- e.g. $a: \{1 \rightarrow 1, 2 \rightarrow 2, 3 \rightarrow 3, 4 \rightarrow 4\}$

Definitions

Word-based Translation Model

- Generative model
 - break up translation process into smaller steps
- Translation probability
 - for a source sentence $f = f_1, \dots, f_F$ of length F
 - to an target sentence $e = e_1, \dots, e_E$ of length E
 - with an alignment of each *target* word e_j to a foreign word f_i according to the alignment function $a: j \rightarrow i$

$$p(e, a | f) = \prod_{j=1}^E t(e_j | f_{a(j)})$$

Recap

- Have parallel corpus
- Need translation rules
- If we have alignments
 - Collect counts
 - Compute translation rules
- If we have translation rules
 - Compute probabilities
 - Implied alignments
- ‘Chicken and Egg’ situation

Algorithm

Initialize *translation rules*

Do until (convergence) {

Translate *parallel corpus*

Count *translations*

Create *translation rules*

}

Example

Parallel Corpus

Das Haus

The house

Das Buch

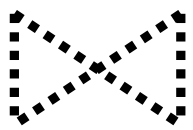
The book

Ein Buch

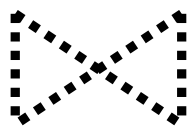
A book

Convergence

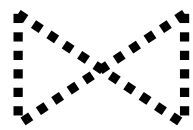
das Haus
the house



das Buch
the book



ein Buch
a book

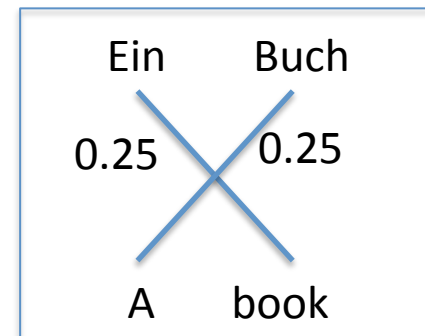
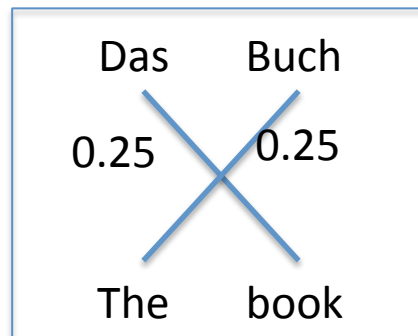
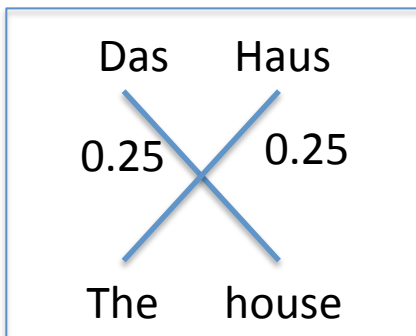
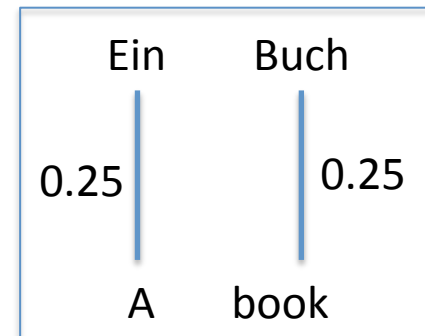
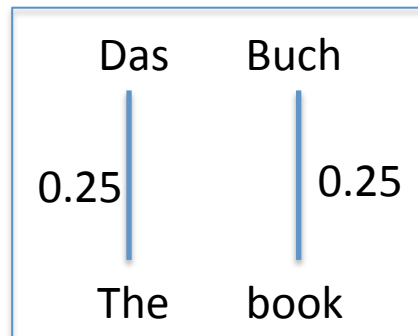
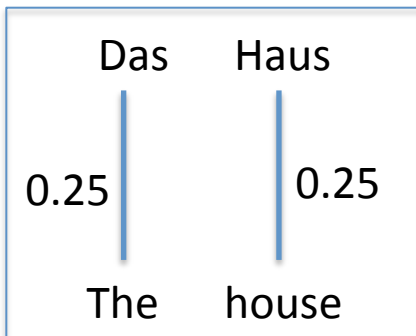


e	f	initial
the	das	0.25
book	das	0.25
house	das	0.25
the	buch	0.25
book	buch	0.25
a	buch	0.25
book	ein	0.25
a	ein	0.25
the	haus	0.25
house	haus	0.25

Example

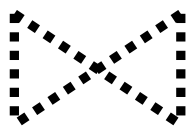
Iteration 1

Calculate probability, for all possible alignments

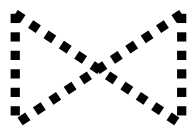


Convergence

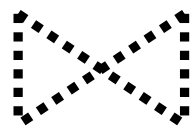
das Haus
the house



das Buch
the book



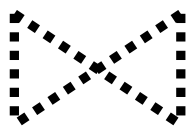
ein Buch
a book



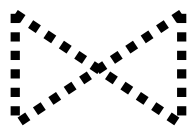
e	f	initial	1st it.
the	das	0.25	0.5
book	das	0.25	0.25
house	das	0.25	0.25
the	buch	0.25	0.25
book	buch	0.25	0.5
a	buch	0.25	0.25
book	ein	0.25	0.5
a	ein	0.25	0.5
the	haus	0.25	0.5
house	haus	0.25	0.5

Convergence

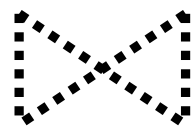
das Haus
the house



das Buch
the book



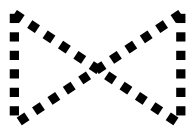
ein Buch
a book



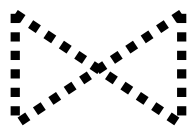
e	f	initial	1st it.	2nd it.
the	das	0.25	0.5	0.6364
book	das	0.25	0.25	0.1818
house	das	0.25	0.25	0.1818
the	buch	0.25	0.25	0.1818
book	buch	0.25	0.5	0.6364
a	buch	0.25	0.25	0.1818
book	ein	0.25	0.5	0.4286
a	ein	0.25	0.5	0.5714
the	haus	0.25	0.5	0.4286
house	haus	0.25	0.5	0.5714

Convergence

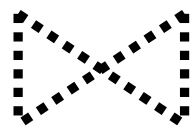
das Haus
the house



das Buch
the book



ein Buch
a book



e	f	initial	1st it.	2nd it.	3rd it.	...	final
the	das	0.25	0.5	0.6364	0.7479	...	1
book	das	0.25	0.25	0.1818	0.1208	...	0
house	das	0.25	0.25	0.1818	0.1313	...	0
the	buch	0.25	0.25	0.1818	0.1208	...	0
book	buch	0.25	0.5	0.6364	0.7479	...	1
a	buch	0.25	0.25	0.1818	0.1313	...	0
book	ein	0.25	0.5	0.4286	0.3466	...	0
a	ein	0.25	0.5	0.5714	0.6534	...	1
the	haus	0.25	0.5	0.4286	0.3466	...	0
house	haus	0.25	0.5	0.5714	0.6534	...	1

EM Algorithm

Expectation-Maximization

Initialize translation model

Repeat until you're bored {

 Expectation step:

 Apply translation model to the data

 Compute 'expected' alignment

 Maximization step:

 Take computed alignment as fact

 Collect counts

 Compute translation model to maximize probability of data

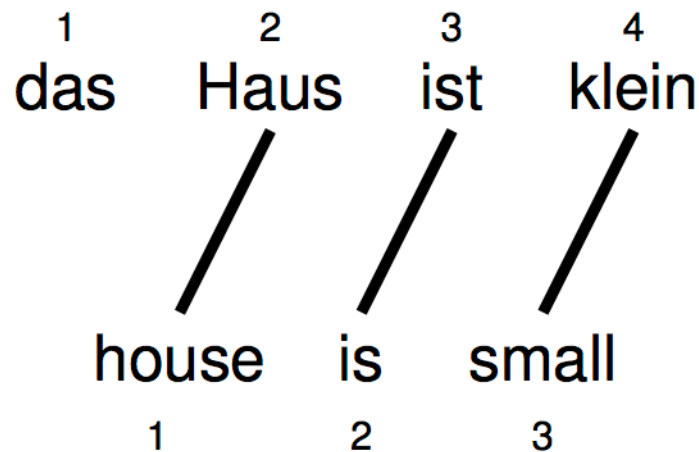
}

Complications

- Lexical translation
- Doesn't say anything about
 1. Dropping words
 2. Inserting
 3. Word order
 4. One-to-many translation

Dropping Words

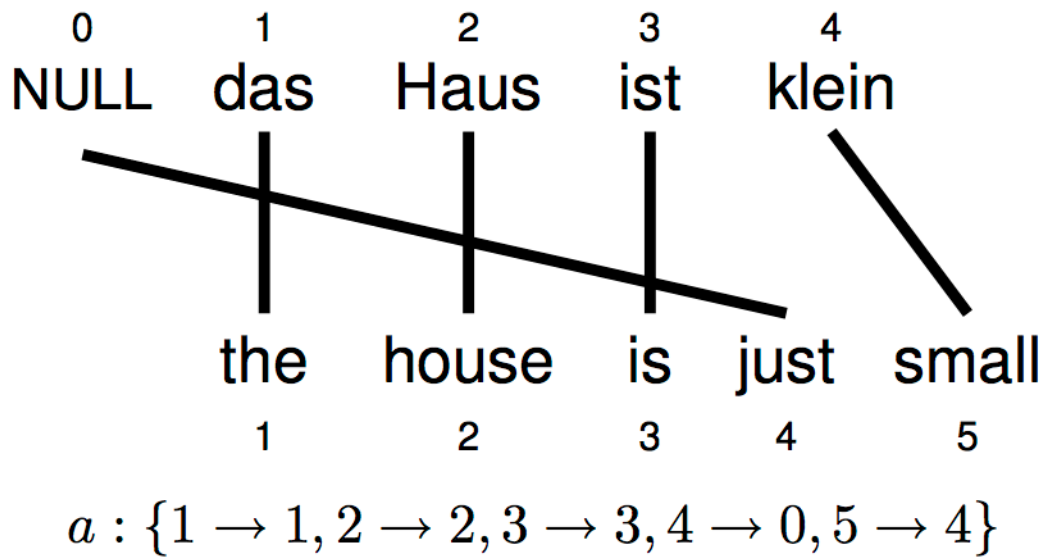
- Words may be dropped when translated
 - German article *das* is dropped



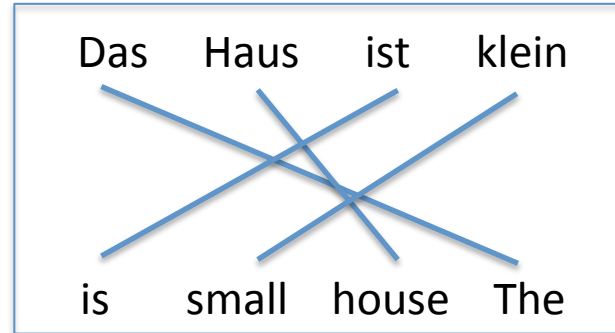
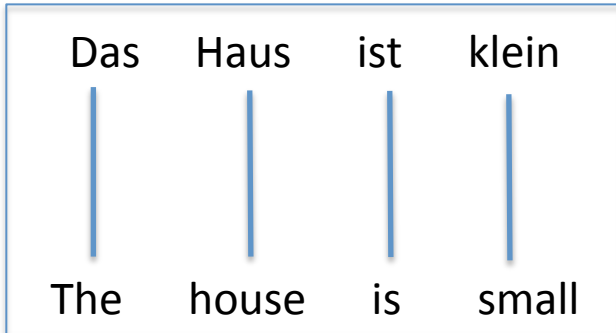
$$a : \{1 \rightarrow 2, 2 \rightarrow 3, 3 \rightarrow 4\}$$

Inserting Words

- Words may be inserted during translation
 - English just does not have German equivalent
 - Align it to source NULL word



Word Order



Same lexical probability

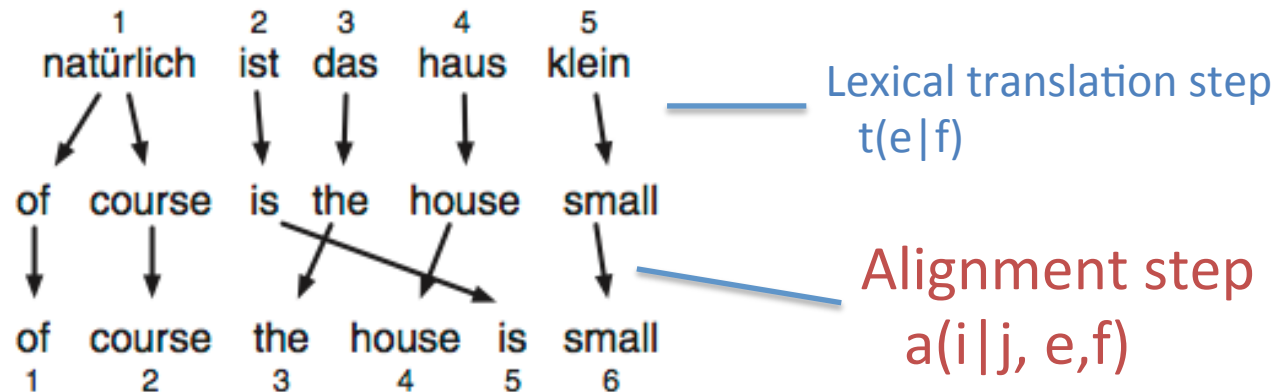
IBM Models

- Model 1 - Lexical probabilities
- Model 2 - Absolute reordering model
- Model 3 - Fertility model
- Model 4 - Relative reordering model
- Model 5- Fixes other deficiencies

- Only Model 1 has global maximum
 - No guarantee of converging to best model for higher models
- Training of higher model builds on previous models
- Computationally biggest change in Model 3
 - exhaustive count collection becomes computationally too expensive
 - sampling over high probability alignments is used instead

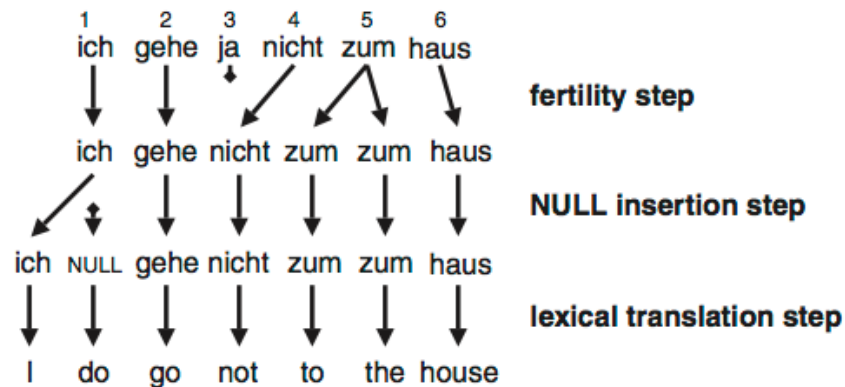
IBM Model 2

- Absolute reordering



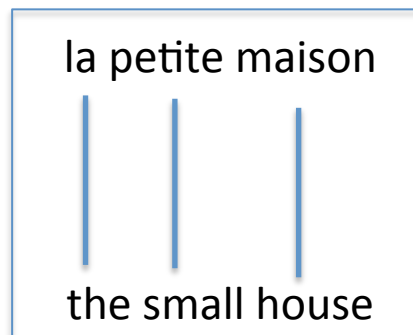
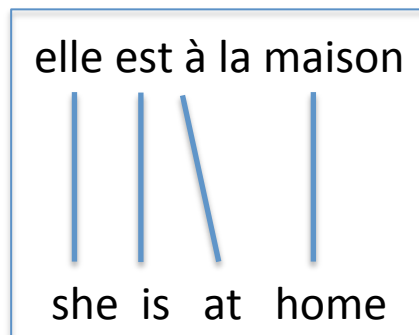
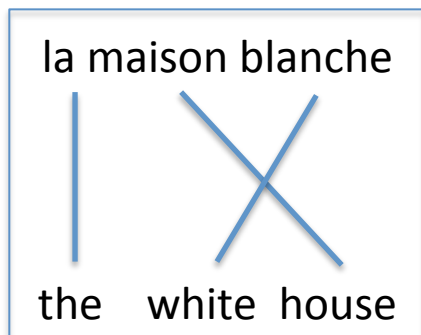
IBM Model 3

- Fertility
 - Probability that a source word create 0, 1, 2... number of target words
 $n(x|f)$
eg. $n(1|\text{Haus}) \approx 1$
 $n(2|\text{Waschmaschine}) \approx 1$
 - Word deletion
eg. $n(0|\text{ja}) \approx 1$
 - Word insertion
 - Special NULL source word
 $n(x|\text{NULL})$



Word-Based Model

Parallel Corpus



Counts

$\text{Count}(\text{la}, \text{the}) = 2$

$\text{Count}(\text{maison}, \text{house}) = 2$

$\text{Count}(\text{maison}, \text{home}) = 1$

$\text{Count}(\text{blanche}, \text{white}) = 1$

$\text{Count}(\text{elle}, \text{she}) = 1$

$\text{Count}(\text{est}, \text{is}) = 1$

$\text{Count}(\text{à}, \text{at}) = 1$

$\text{Count}(\text{petite}, \text{small}) = 1$

Translation Rules

$t(\text{the}, \text{la}) = 1.0$

$t(\text{house} | \text{maison}) = 0.33333$

$t(\text{home} | \text{maison}) = 0.66666$

$t(\text{white} | \text{blanche}) = 1.0$

$t(\text{she} | \text{elle}) = 1.0$

$t(\text{is} | \text{est}) = 1.0$

$t(\text{at} | \text{à}) = 1.0$

$t(\text{small} | \text{petite}) = 1.0$

Phrase-Based Model

Word-Based Model

$$t(\text{the}, \text{la}) = 1.0$$

$$t(\text{house} | \text{maison}) = 0.33333$$

$$t(\text{home} | \text{maison}) = 0.66666$$

$$t(\text{she} | \text{elle}) = 1$$

$$t(\text{is} | \text{est}) = 1$$

Phrase-Based Model

$$t(\text{the}, \text{la}) = 1.0$$

$$t(\text{house} | \text{maison}) = 0.33333$$

$$t(\text{home} | \text{maison}) = 0.66666$$

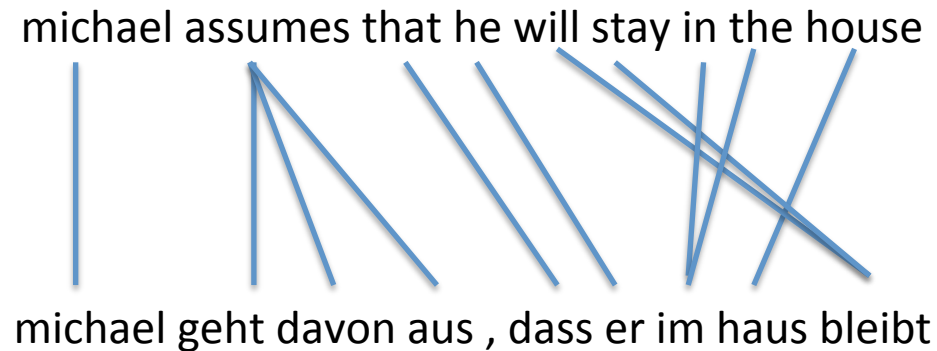
$$t(\text{the house} | \text{la maison}) = ???$$

$$t(\text{the home} | \text{la maison}) = ???$$

$$t(\text{she is} | \text{elle est}) = ???$$

Phrase Extraction

Word Alignment



Grid Representation

	michael	geht	davon	aus	,	dass	er	im	haus	bleibt
michael										
assumes										
that										
he										
will										
stay										
in										
the										
house										

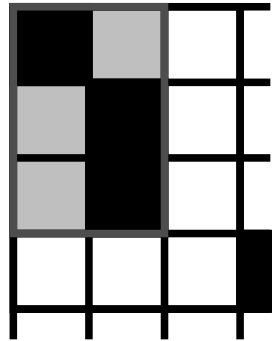
Phrase Extraction

	michael	geht	davon	aus	,	dass	er	im	haus	bleibt
michael										
assumes										
that										
he										
will										
stay										
in										
the										
house										

Extract phrase-pair consistent with word alignment

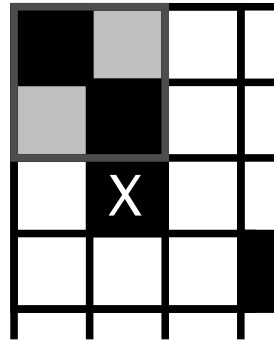
assumes that → geht davon aus , das

Consistent



consistent

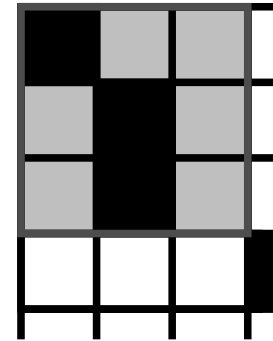
ok



inconsistent

violated

One alignment
point outside



consistent

ok

Unaligned word
is fine

All words of the phrase-pair have to align to each other
- AND no other word

Phrase Extraction

	michael	geht	davon	aus	,	dass	er	im	haus	bleibt
michael										
assumes										
that										
he										
will										
stay										
in										
the										
house										

Smallest phrase-pair

Michael → michael

Assumes → geht davon aus

Assumes → geht davon aus ,

That → das

That → das ,

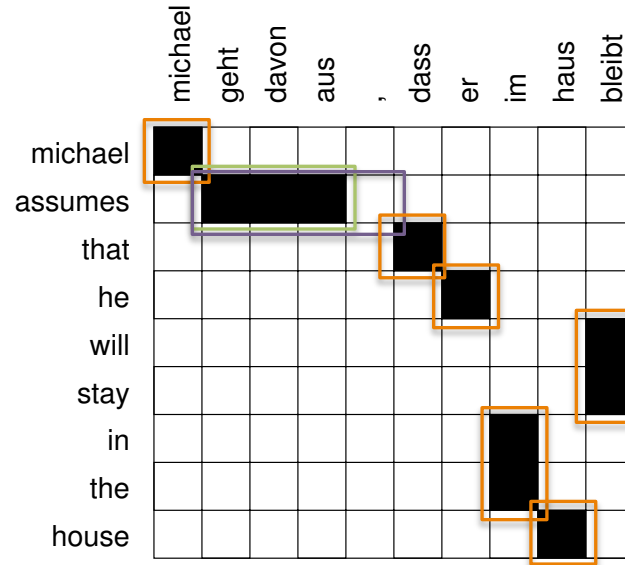
He → er

Will stay → bleibt

In the → im

House → haus

Phrase Extraction



Smallest phrase-pairs

michael → michael

assumes → geht davon aus

assumes → geht davon aus ,

that → das

that → das ,

he → er

will stay → bleibt

In the → im

house → haus

Phrase Extraction

	michael	geht	davon	aus	,	dass	er	im	haus	bleibt
michael	■									
assumes		■	■	■						
that						■				
he							■			
will										■
stay										■
in								■		
the								■		
house									■	

Larger phrase-pairs

michael assume → michael geht davon aus

michael assume → michael geht davon aus ,

assumes that → geht davon aus , dass

That he → das er

That he → , das er

michael assume that → michael geht davon aus , dass

michael assume that he → michael geht davon aus , dass er

.....

Scoring Phrase-Pairs

- Phrase-pair extraction
 - Collect all phrase-pairs from the data
- Phrase-pair scoring
 - Calculate probabilities

$$t(e | f) = \frac{\text{count}(e, f)}{\sum_i \text{count}(e_i, f)}$$

Evaluation

Evaluation

- How good is a given MT system?
- Hard problem
 - Many different acceptable translations
 - Semantic equivalence/similarity
- Evaluation metrics
 - Subjective judgment by human evaluators
 - Automatic metrics
 - Task-based evaluation
 - How much post-editing effort?
 - Does the information come across?

Ten Translations of a Chinese Sentence

这个 机场 的 安全 工作 由 以色列 方面 负责 。

Israeli officials are responsible for airport security.

Israel is in charge of the security at this airport.

The security work for this airport is the responsibility of the Israel government.

Israeli side was in charge of the security of this airport.

Israel is responsible for the airport's security.

Israel is responsible for safety work at this airport.

Israel presides over the security of the airport.

Israel took charge of the airport security.

The safety of this airport is taken charge of by Israel.

This airport's security is the responsibility of the Israeli security officials.

(a typical example from the 2001 NIST evaluation set)

Adequacy and Fluency

- Human judgment
 - Given
 - MT output
 - Source and/or reference translation
 - Assess quality of MT output
- Metrics
 - Adequacy
 - Does the output convey the same meaning as the input sentence?
 - Is part of the message lost, added, or distorted?
 - Fluency
 - Is the output good fluent English?
 - This involves both grammatical correctness and idiomatic word choices.

Adequacy and Fluency Scales

Adequacy	
5	All meaning
4	Most meaning
3	Much meaning
2	Little meaning
1	None

Fluent	
5	Flawless English
4	Good English
3	Non-native English
2	Dis-fluent English
1	Incomprehensible

Automatic Metrics

Goals

- Low Cost
 - Time and money spent on evaluation
- Tunable
 - Automatically optimize system performance towards metric
- Meaningful
 - intuitive interpretation of translation quality
- Consistent
 - repeated use of metric should give same results
- Correct
 - Higher score → better system

Automatic Metrics

- Basic strategy
 - given: machine translation output
 - given: human reference translation
 - task: compute similarity between them

Precision and Recall

SYSTEM A:

Israeli officials ~~responsibility of~~ airport ~~safety~~

REFERENCE:

Israeli officials are responsible for airport security

- Precision

$$\frac{\text{correct}}{\text{output-length}} = \frac{3}{6} = 50\%$$

- Recall

$$\frac{\text{correct}}{\text{reference-length}} = \frac{3}{7} = 43\%$$

- F-measure

$$\frac{\text{precision} \times \text{recall}}{(\text{precision} + \text{recall}) / 2} = \frac{0.5 \times 0.43}{(0.5 + 0.43) / 2} = 46\%$$

Precision and Recall



Metric	System A	System B
Precision	50%	100%
Recall	43%	100%
F-measure	46%	100%

Flaw: no penalty for re-ordering

Word Error Rate

- Minimum number of editing steps to transform output to reference
 - **Match:** words match, no cost
 - **Substitution:** replace one word with another
 - **Insertion:** add word
 - **Deletion:** drop word
- Levenshtein distance

$$WER = \frac{\text{substitutions} + \text{insertions} + \text{deletions}}{\text{reference-length}}$$

Example

		Israeli	officials	responsibility	of	airport	safety
	0	1	2	3	4	5	6
Israeli	1	0	1	2	3	4	5
officials	2	1	0	1	2	3	4
are	3	2	1	1	2	3	4
responsible	4	3	2	2	2	3	4
for	5	4	3	3	3	3	4
airport	6	5	4	4	4	3	4
security	7	6	5	5	5	4	4

		airport	security	Israeli	officials	are	responsible
	0	1	2	3	4	5	6
Israeli	1	1	2	2	3	4	5
officials	2	2	2	3	2	3	4
are	3	3	3	3	3	2	3
responsible	4	4	4	4	4	3	2
for	5	5	5	5	5	4	3
airport	6	5	6	6	6	5	4
security	7	6	5	6	7	6	5

Metric	System A	System B
Word Error Rate	57%	71%

BLEU

- N-gram overlap between
 - MT output
 - Reference
- Compute precision for n-gram of size 1 to 4
- Brevity penalty (penalize short translations)
- Over entire test set, not individual sentences

$$BLEU = \min\left(1, \frac{\text{output-length}}{\text{reference-length}}\right) \left(\prod_{i=1}^4 precision_i\right)^{1/4}$$

BLEU

SYSTEM A: Israeli officials responsibility of airport safety
2-GRAM MATCH 1-GRAM MATCH

REFERENCE: Israeli officials are responsible for airport security

SYSTEM B: airport security Israeli officials are responsible
2-GRAM MATCH 4-GRAM MATCH

Metric	System A	System B
1-gram precision	3/6	6/6
2-gram precision	1/5	4/5
3-gram precision	0/4	2/4
4-gram precision	0/3	1/3
Brevity penalty	6/7	6/7
BLEU	0%	52%

Multiple References

SYSTEM: Israeli officials responsibility of airport safety
2-GRAM MATCH 2-GRAM MATCH 1-GRAM

REFERENCES: Israeli officials are responsible for airport security
Israel is in charge of the security at this airport
The security work for this airport is the responsibility of the Israel government
Israeli side was in charge of the security of this airport

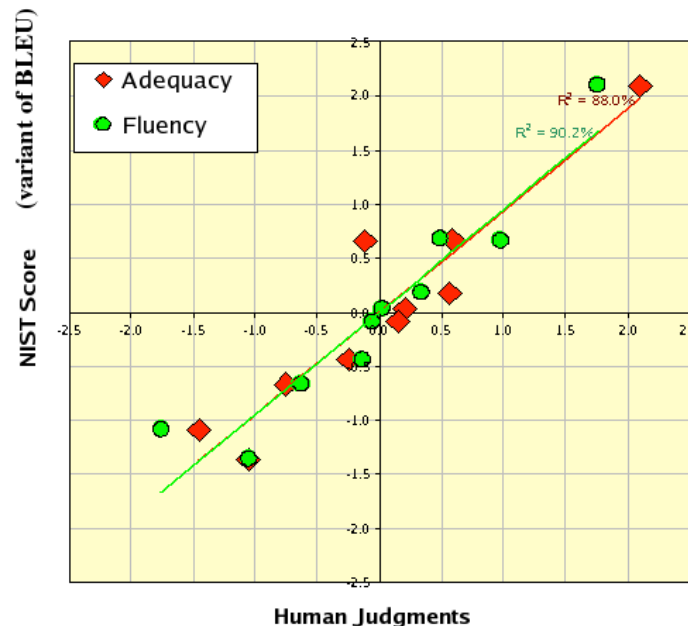
- Account for variability
 - Use multiple reference translations
 - N-grams may match ANY of the references
 - Closest reference length used

Critique of Automatic Metrics

- Ignore relevance of words
 - names and core concepts more important than determiners and punctuation
- Operate on local level
 - do not consider overall grammaticality of the sentence or sentence meaning
- Scores are meaningless
 - scores very test-set specific, absolute value not informative
- Human translators score low on BLEU
 - possibly because of higher variability, different word choices

Evaluation of Evaluation Metrics

- Automatic metrics are low cost, tunable, consistent
 - But are they correct?
 - Yes, if they correlate with human judgement



(Doddington, 2002)?

Metric Research

- Syntactic Similarity
- Semantic equivalence or entailment
- Metric targeted at reordering
- Training metric
- etc

Next Time

- Read Zens et al. (2002)
- Come with your MT team spirit