CA446

Statistical Machine Translation

Week 6: IBM Models

Lecturer: Qun Liu

Lab Tutor: Xiaofeng Wu, Iacer Calixto

2nd Semester, 2014-2015 Academic Year

http://computing.dcu.ie/~qliu/CA446

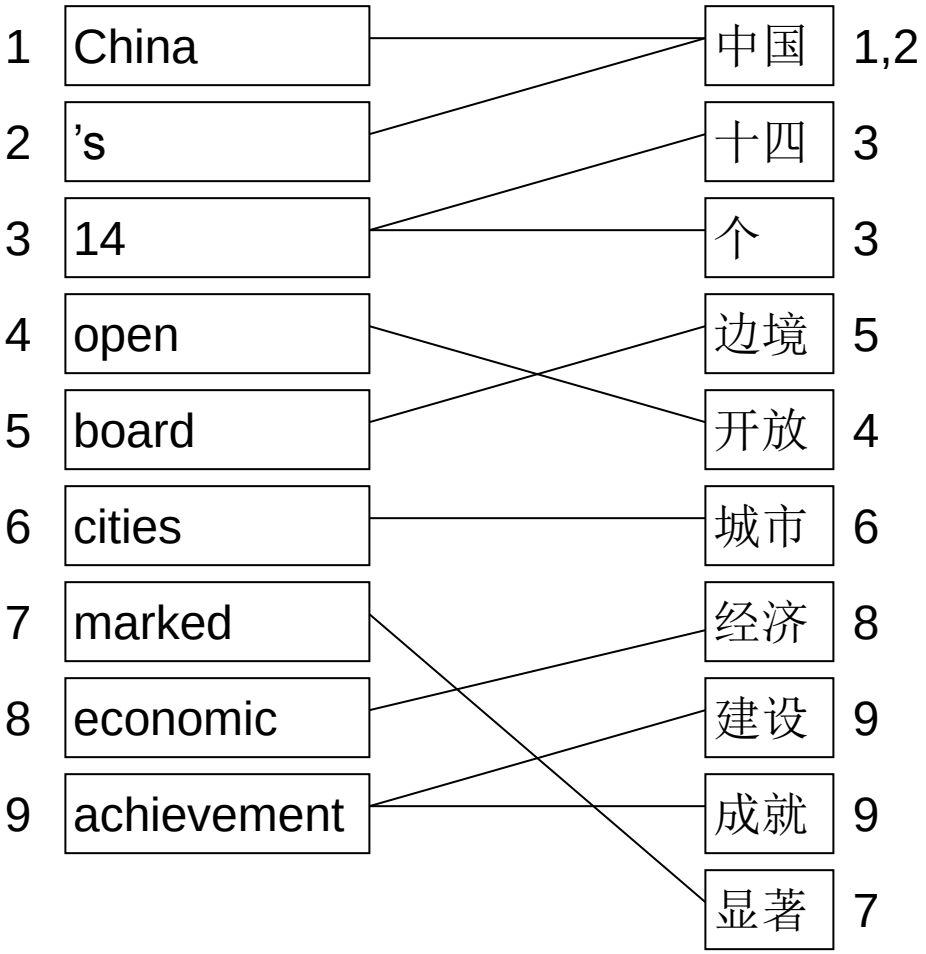
Content

- **IBM Model 1**
- Higher IBM Models
- Project and Homework
- Exercises

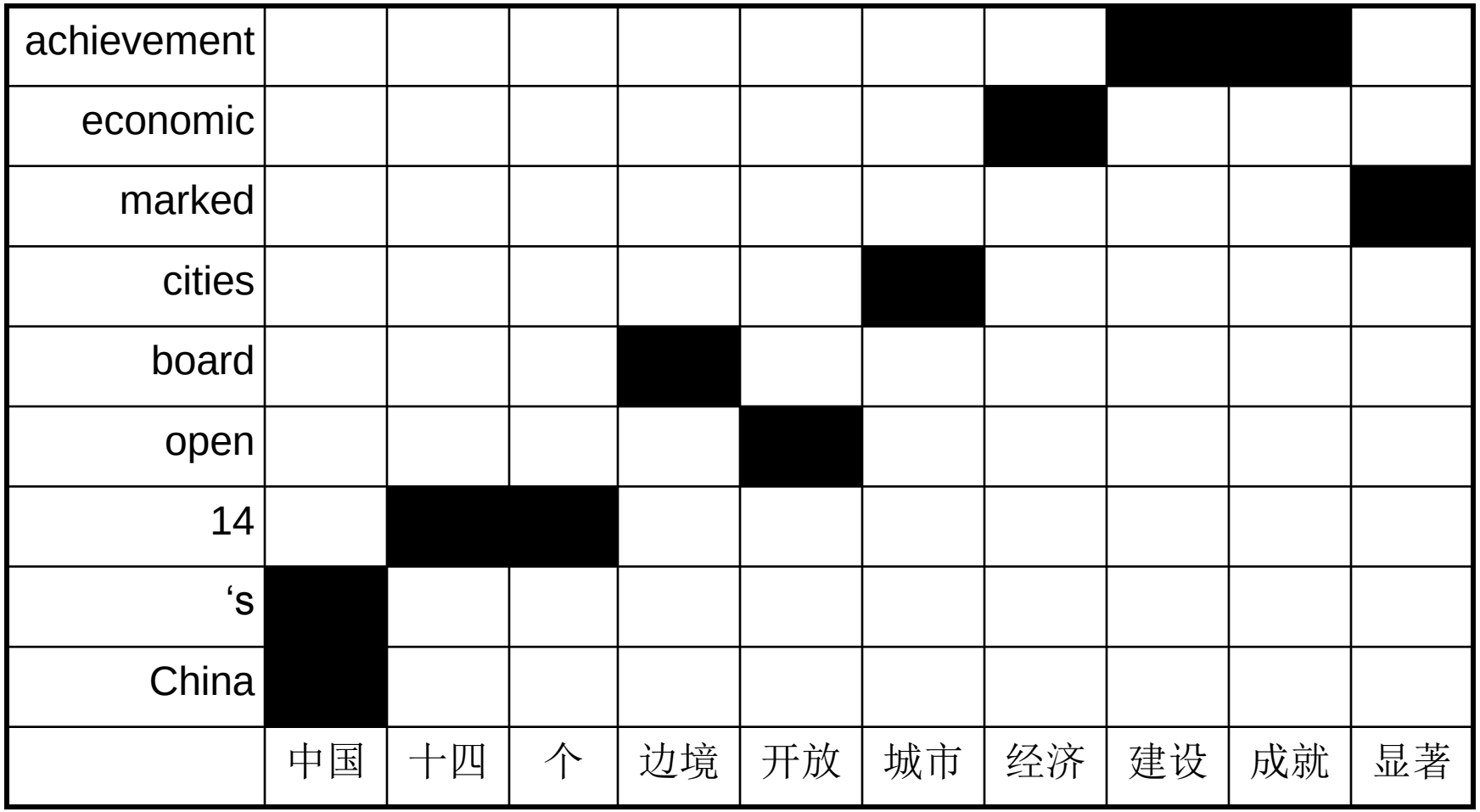
Translation Model & Word Alignment

- Translation Model: $p(F|E)$
- It is impossible to estimate the translation model directly.
- We break $p(F|E)$ down into multiplication of the probability of the word translation probabilities $t(f|e)$
- We need to define an alignment between words to implement this breakdown

Word Alignment



Word Alignment



Translation Model & Word Alignment

- Consider all the possible alignment of a pair of sentence:

$$p(F|E) = \sum_{A} \underbrace{p(F, A|E)}$$

The probability of
a word alignment A
given source sentence E

Word Alignment in IBM Models

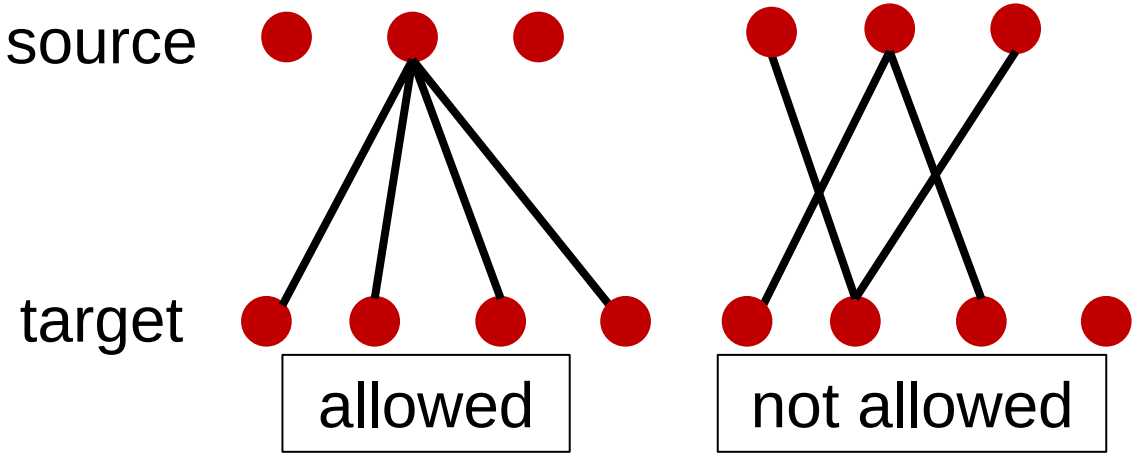
- We express the word alignment as:

$$A = a_1^m = a_1 a_2 \dots a_m$$
$$\forall i \in \{1, \dots, m\}, a_i \in \{0, \dots, l\}$$

$a_1 = 1 \quad a_2 = 2 \quad a_3 = 3 \quad a_4 = 0 \quad a_5 = 4$

Word Alignment in IBM Models

- Limitations:
  - Much relaxed than the limitation we used in the examples for EM algorithm where only one-to-one alignments are allowed
  - Still limited because many-to-one alignments are still not allowed



Assumption of IBM Model 1

- We assume that:
  - $p(F, A|E)$ is independent of the length of the target sentence m
  - $p(F, A|E)$ is independent of the target word order
- Then we have:

$$p(F, A|E) = \prod_{j=1}^{m} t(f_j|e_{a_j})$$

IBM Model 1

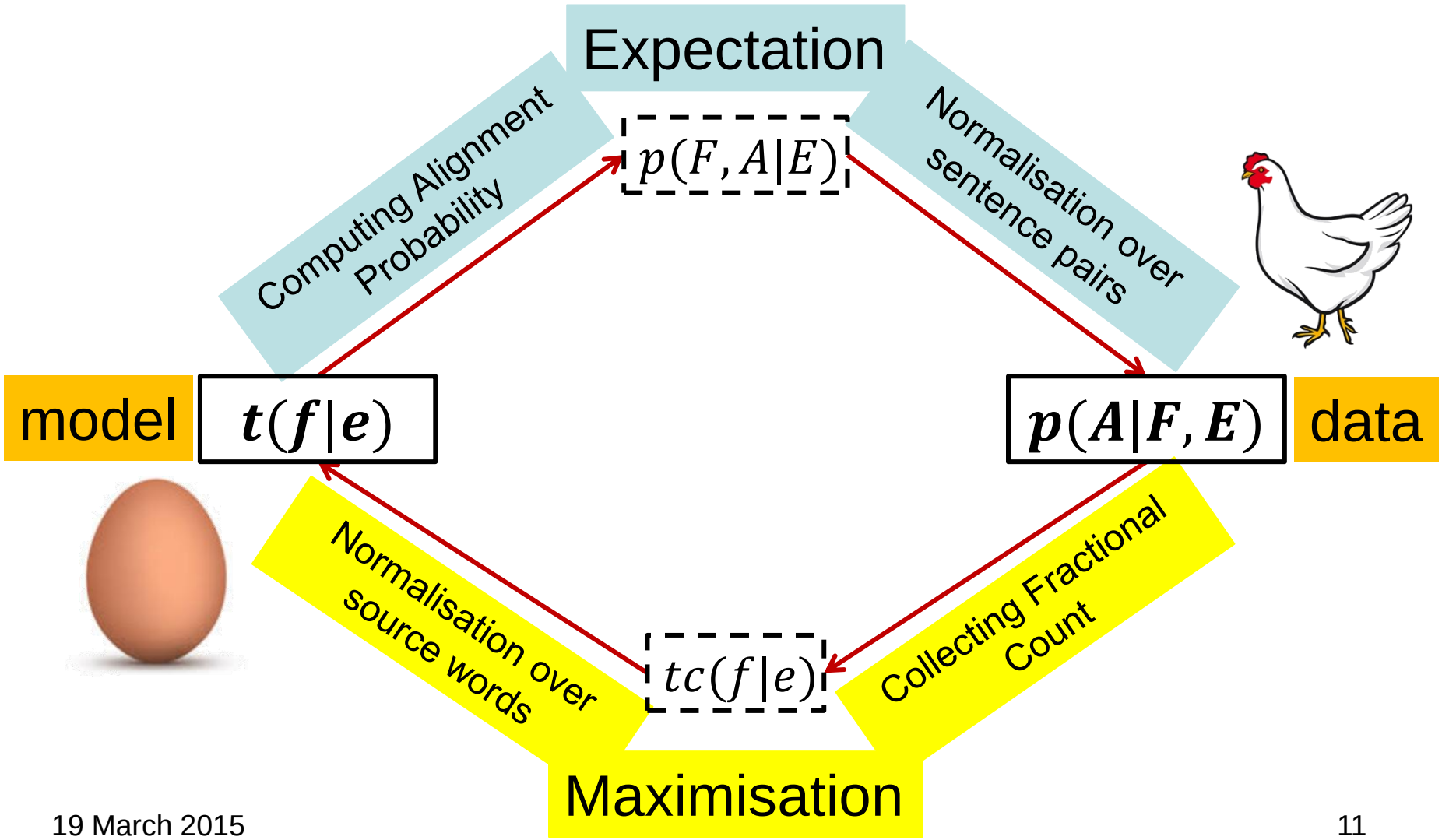
$$p(F|E) = \sum_{A} p(F, A|E) = \sum_{A} \prod_{j=1}^{m} t(f_j|e_{a_j})$$

Because: $\forall i \in \{1, \ldots, m\}, a_i \in \{0, \ldots, l\}$

$$p(F|E) = \underbrace{\sum_{a_{1=0}}^{l} \sum_{a_{2=0}}^{l} \ldots \sum_{a_{m=0}}^{l} \prod_{j=1}^{m}}_{} t(f_j|e_{a_j})$$

Complexity: $(l+1)^m \times m$: intractable

EM algorithm



EM algorithm

$$p(A|F,E) = \frac{p(A,F|E)}{p(F|E)} = \frac{p(A,F|E)}{\sum_A p(A,F|E)}$$

- In EM algorithm, we need to calculate the sum of the probabilities of possible word alignments of a specific sentence pair.

Sum on all Alignments

- $t(f_1|e_1) \times t(f_2|e_1) \times t(f_3|e_1) \times \cdots \times t\left(f_{l_f}\middle|e_1\right) +$
- $t(f_1|e_1) \times t(f_2|e_1) \times t(f_3|e_1) \times \cdots \times t\left(f_{l_f}\middle|e_2\right) +$
- $t(f_1|e_1) \times t(f_2|e_1) \times t(f_3|e_1) \times \cdots \times t\left(f_{l_f}\middle|e_3\right) +$
- ...
- $t(f_1|e_1) \times t(f_2|e_2) \times t(f_3|e_1) \times \cdots \times t\left(f_{l_f}\middle|e_1\right) +$
- $t(f_1|e_1) \times t(f_2|e_2) \times t(f_3|e_1) \times \cdots \times t\left(f_{l_f}\middle|e_2\right) +$
- $t(f_1|e_1) \times t(f_2|e_1) \times t(f_3|e_1) \times \cdots \times t\left(f_{l_f}\middle|e_3\right) +$
- ...
- $t(f_1|e_2) \times t(f_2|e_2) \times t(f_3|e_2) \times \cdots \times t\left(f_{l_f}\middle|e_1\right) +$
- $t(f_1|e_2) \times t(f_2|e_2) \times t(f_3|e_2) \times \cdots \times t\left(f_{l_f}\middle|e_2\right) +$
- $t(f_1|e_2) \times t(f_2|e_2) \times t(f_3|e_2) \times \cdots \times t\left(f_{l_f}\middle|e_3\right) +$
- ...

Sum on all Alignments

- However, there are regularities in this expression which we can factor out.
  - For example, we can factor out $t(f_1|e_1)$ from all the rows it occurs in.
  - Difference between $xy + xz$ (3 arithmetic operations) and $x(y + z)$ (2 arithmetic operations)

Sum on all Alignments

$$t(f_1|e_1) \times \begin{bmatrix} t(f_2|e_1) \times t(f_3|e_1) \times \cdots \times t\left(f_{l_f}\middle|e_1\right) \\ t(f_2|e_1) \times t(f_3|e_1) \times \cdots \times t\left(f_{l_f}\middle|e_2\right) \\ \cdots \\ t(f_2|e_2) \times t(f_3|e_1) \times \cdots \times t\left(f_{l_f}\middle|e_1\right) \\ t(f_2|e_2) \times t(f_3|e_1) \times \cdots \times t\left(f_{l_f}\middle|e_2\right) \\ t(f_2|e_1) \times t(f_3|e_1) \times \cdots \times t\left(f_{l_f}\middle|e_3\right) \\ \cdots \end{bmatrix} +$$

$$\cdots$$

$$t(f_1|e_2) \times t(f_2|e_2) \times t(f_3|e_2) \times \cdots \times t\left(f_{l_f}\middle|e_1\right) +$$

$$t(f_1|e_2) \times t(f_2|e_2) \times t(f_3|e_2) \times \cdots \times t\left(f_{l_f}\middle|e_2\right) +$$

$$t(f_1|e_2) \times t(f_2|e_2) \times t(f_3|e_2) \times \cdots \times t\left(f_{l_f}\middle|e_3\right) +$$

$$\cdots$$

$$t(f_1|e_1) \times t(f_2|e_1) \times t(f_3|e_1) \times \cdots \times t\left(f_{l_f}\middle|e_{l_e}\right) +$$

Sum on all Alignments

- However, there are regularities in this expression which we can factor out.
  - For example, we can factor out $t(f_1|e_1)$ from all the rows it occurs in.
  - Difference between $xy + xz$ (3 arithmetic operations) and $x(y+z)$ (2 arithmetic operations)
- Factoring out continually gives us:

$$p(F|E) = \sum_{a_{1=0}}^{l} \sum_{a_{2=0}}^{l} \dots \sum_{a_{m=0}}^{l} \prod_{j=1}^{m} t\left(f_j \middle| e_{a_j}\right) = \prod_{j=1}^{m} \sum_{i=0}^{l} t(f_j|e_i)$$

Complexity: $(l+1) \times m$: tractable

IBM Model 1

Iterative formula:

$$t(f|e) = \lambda_e^{-1} tc(f|e;F,E)$$

$$tc(f|e;F,E) = \frac{t(f|e)}{t(f|e_0) + \cdots + t(f|e_l)} \underbrace{\sum_{j=1}^{m} \delta(f, f_j)}_{\#\text{ of } f \text{ in } F} \underbrace{\sum_{i=1}^{l} \delta(e, e_i)}_{\#\text{ of } e \text{ in } E}$$

Example: Factoring Out

Given the sentence pair:

the house – la maison

- calculate $p(a, f|e)$ for all alignments and then
- perform the factoring out trick.

Example: Factoring Out

- $p(a, f|e)$ for all alignments:

$$\sum_{a_{1=0}}^{l} \sum_{a_{2=0}}^{l} \dots \sum_{a_{m=0}}^{l} \prod_{j=1}^{m} t\left(f_j \middle| e_{a_j}\right)$$

To simplify the problem, we assume $a_j \neq 0$

Example: Factoring Out

$$t(maison|house) * t(la|the)$$
$$+$$
$$t(maison|the) * t(la|house)$$
$$+$$
$$t(maison|the) * \ t(la|the)$$
$$+$$
$$t(maison|house) * t(la|house)$$

Example: Factoring Out

$$t(maison|house) * (t(la|the) + t(la|house))$$
$$+$$
$$t(maison|the) * t(la|house)$$
$$+$$
$$t(maison|the) * t(la|the)$$

Example: Factoring Out

$$t(maison|house) * (t(la|the) + t(la|house)) + t(maison|the) * (t(la|the) + t(la|house))$$

Example: Factoring Out

$$(t(maison|house) + t(maison|the))$$
$$*$$
$$(t(la|the) + t(la|house))$$

Example: Factoring Out

- $p(a, f|e)$ for all alignments:

$$\sum_{a_{1=0}}^{l}\sum_{a_{2=0}}^{l}\dots\sum_{a_{m=0}}^{l}\prod_{j=1}^{m}t\left(f_j\middle|e_{a_j}\right)=\prod_{j=1}^{m}\sum_{i=0}^{l}t\left(f_j|e_i\right)$$

Content

- IBM Model 1
- **Higher IBM Models**
- Project and Homework
- Exercises

IBM Model 1

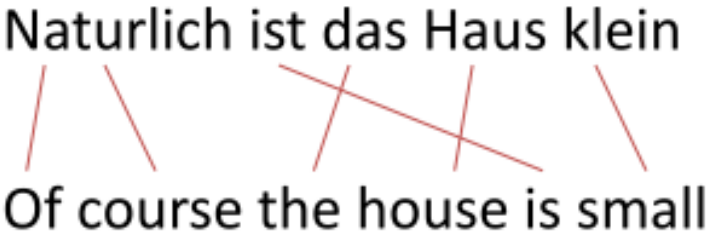
Lexical / word-to-word translation parameters

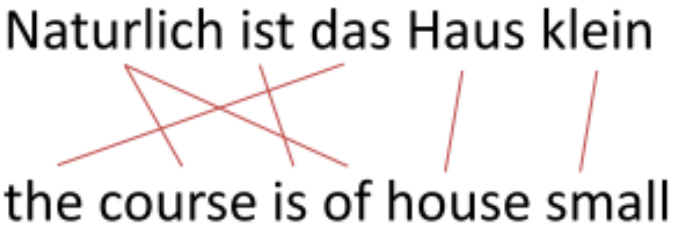
$$t(house|Haus)$$
$$t(building|Haus)\ldots$$

Note: Assume for all of these examples that we are calculating the probability of an English sentence given a German sentence.

Deficiencies of IBM Model 1

With IBM Model 1, both the following translations would be given the same probability:





Higher IBM Models

Higher IBM models take more information into account:

- n: Fertility parameters
- $n(1|klitzklein)$
- $n(2|klitzklein)$ …
- i.e. what is the probability that “klitzklein” will produce exactly 1 or 2 English words?

Higher IBM Models

d: Distortion parameters

- $d(2|2)$
- $d(3|2)$ …
- i.e. what is the probability that the German word in position 2 of the German sentence will generate an English word that ends up in position 2/3 of an English translation?

Higher IBM Models

Enhanced distortion scheme takes into account the lengths of the German and English sentences:

- $d(3|2,4,6)$: Same as for d(3|2), except we also specify that the given German string has 4 words and the given English string has 6 words

Higher IBM Models

We also have word-translation parameters corresponding to insertions:

- $t(just \mid NULL) = ?$
- i.e. what is the probability that the English word just is inserted ?

Higher IBM Models

Type of information	Probability Distribution Name	Description
Lexical Translation	t	Table plotting source words against target words
Fertility	n	Table plotting source words against fertilities
Distortion	d	Table plotting sentence positions in source against sentence positions in target
Insertion	$p1$	Single number indicating the probability of insertion

Summary of IBM Models

- IBM Model 1 - lexical translation
- IBM Model 2 - adds absolute reordering model
- IBM Model 3 - adds fertility model
- IBM Model 4 - relative reordering model

Summary of IBM Models

- IBM Model 1 - lexical translation
- IBM Model 2 - adds absolute reordering model
- IBM Model 3 - adds fertility model
- IBM Model 4 - relative reordering model

Absolute versus relative reordering

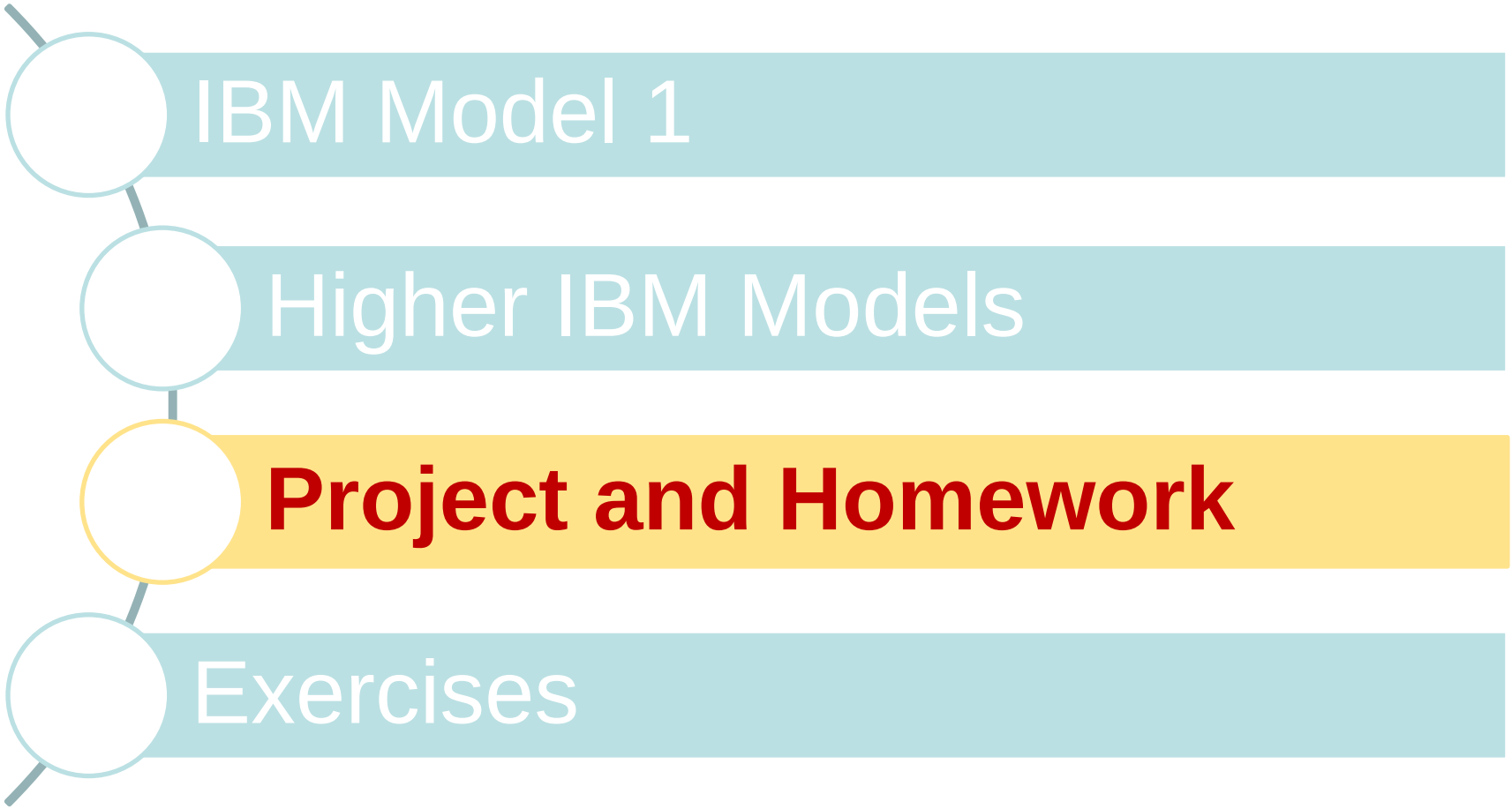
- IBM Model 4 takes into account the fact that words do not move independently of one another.

- The probability of a word moving is conditioned on the probability of the previous word moving.

Parameter Estimation for higher IBM models

- Increased sophistication of higher IBM models means that the computational trick no longer works for summing over all possible alignments
  - Exhaustive count collection becomes computationally too expensive
  - Instead, we have to sum over high probability alignments

Content


IBM Model 1
Higher IBM Models
Project and Homework
Exercises

Schedule of Next Half Semester

- Lectures:
  - Week 7-9: Lectures
  - Week 10: Review and the Scope of Exam
  - Week 11: Project Reports
  - Week 12: PhD Presentations

Schedule of Next Half Semester

- Project: (30%)
  - Individual or Group Work (2-3 members)
  - Research Report:
    - Submission Deadline: Week 11 (Monday)
  - Oral Presentation:
    - Time: Week 11 (Thursday, Friday)

Schedule of Next Half Semester

- Homework: (20%)
  - Individual work
  - Assignment #1: (10%)
    - Topic: Language Model
    - Releasing: Week 6 (Friday)
    - Submission Deadline: week 7 (Friday)
  - Assignment #2: (10%)
    - Topic: Translation Model
    - Releasing: Week 7 (Friday)
    - Submission Deadline: week 8 (Friday)

Content

- IBM Model 1
- Higher IBM Models
- Project and Homework
- **Exercises**

Exercise 1

Given a sentence aligned corpus

{das Haus – the house, das Buch – the book, ein Buch – a book},

calculate the lexical probabilities of

$t(the|das)$
$t(house|das)$
$t(book|das)$
$t(a|das)$

after the first iteration of EM.

Exercise 2

Given the sentence pair *the house – la maison*, calculate p(a,f|e) for all alignments and then perform the factoring out trick.

Exercise 3

Provide examples where unigram, bigram, trigram and 4-gram language models would fail to capture a grammatical constraint of the English language.

Discussion

Acknowledgement

Parts of the content of this lecture are taken from previous lectures and presentations given by Jennifer Foster, Declan Groves, Yvette Graham, Kevin Knight, Josef van Genabith, Andy Way.