COMP90042
Web search and text analysis

Workshop Week 12

xudong.han@unimelb.edu.au
https://github.com/HanXudong/COMP90042_Workshops

Q1
What aspects of human language make automatic translation difficult?

- 南京市　　长江大桥
  Nanjing　　Yangtze River Bridge

- 南京　　市长　　江大桥
  Daqiao Jiang, the major of Nanjing City.

- Not just simple word for word translation
  - structural changes
  - multiple word translations
  - inflections for gender
  - missing information

Q2

- Representation:

  E = $e_1 \ldots e_I$ = *And the program has been implemented*

  F = $f_1 \ldots f_J$ = *Le programme a ete mis en application*

  A = $a_1 \ldots a_J$ = 2, 3, 4, 5, 6, 6, 6.

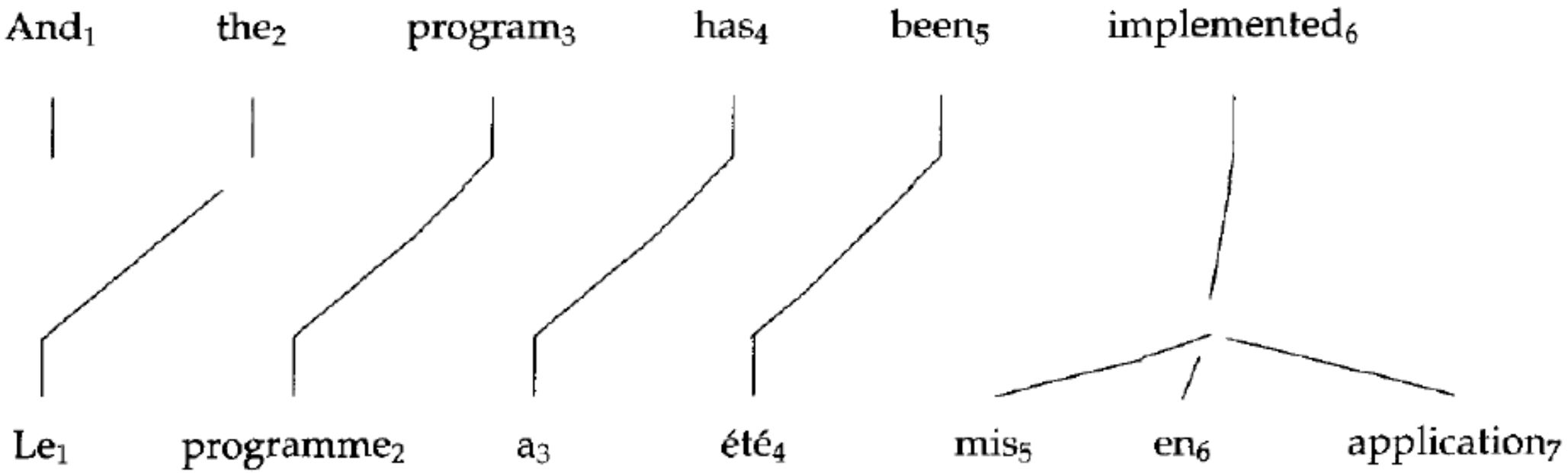


Figure from Brown, Della Pietra, Della Pietra, Mercer, 1993

Q2

- Two components:

Translation Model (TM)

$$\hat{e} = \text{argmax}_e \, P(e)P(f|e)$$

Language Model (LM)

- Responsible for:
  - *P(f|e)* rewards good translations, but permissive of disfluent *e*
  - *P(e)* rewards *e* which look like fluent English, and helps put words in the correct order

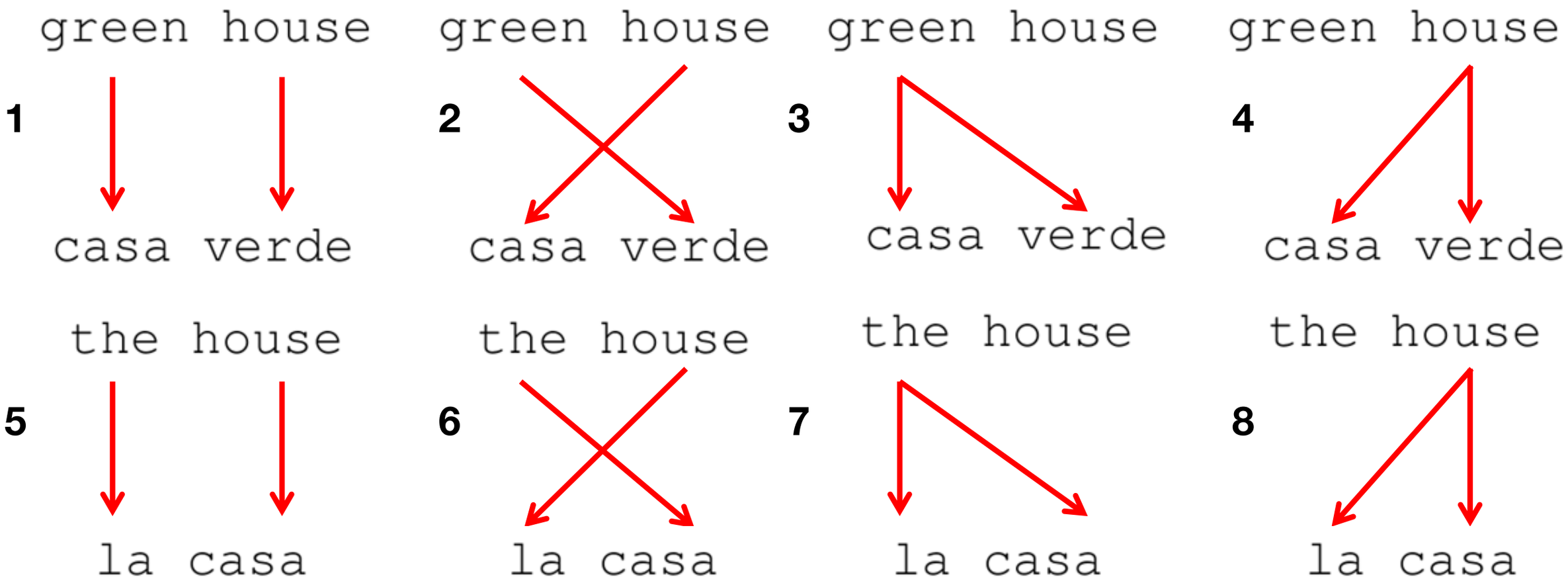
Translate B -> A

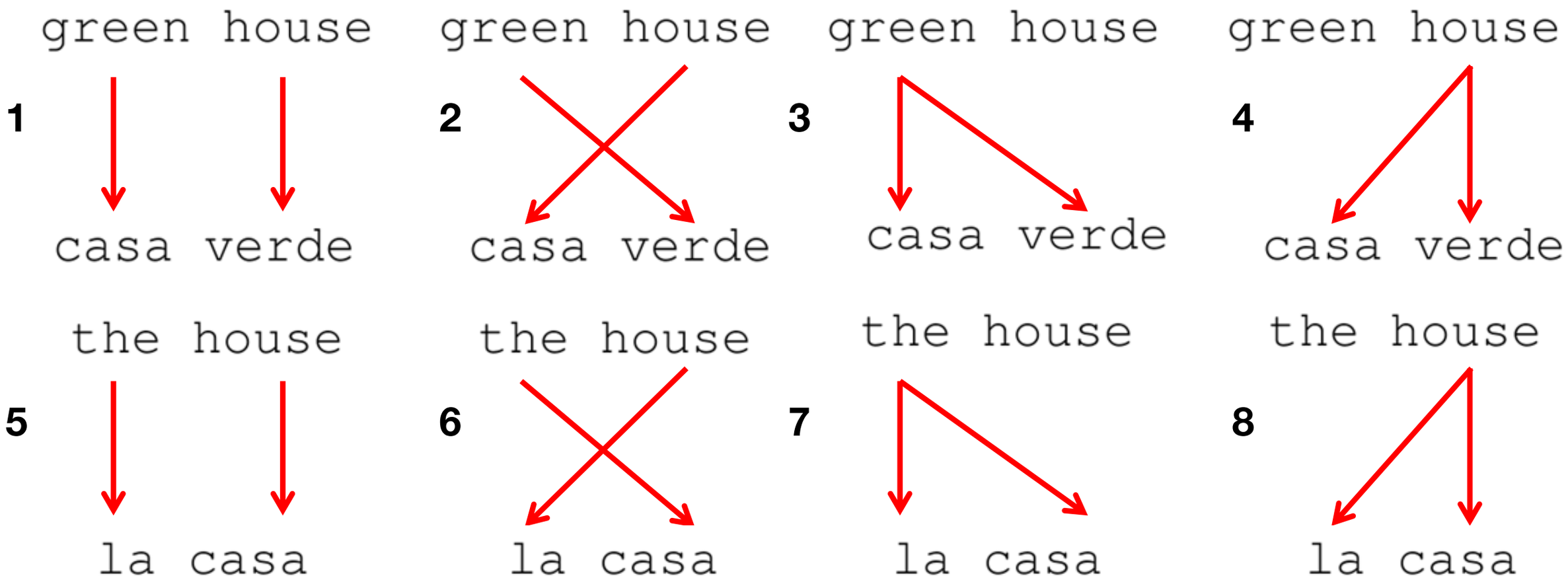
Language A	Language B
green house	casa verde
the house	la casa

t(B|A)	casa	la	verde	Total
green	1/3	1/3	1/3	1
house	1/3	1/3	1/3	1
the	1/3	1/3	1/3	1

- Need to calculate expected alignments under the model (step 2)

$$P(\mathbf{a}|\mathbf{e},\mathbf{f}) = \frac{P(\mathbf{f},\mathbf{a},\mathbf{e})}{P(\mathbf{f},\mathbf{e})} = \frac{P(\mathbf{f},\mathbf{a}|\mathbf{e})}{P(\mathbf{f}|\mathbf{e})}$$





t(B|A)	casa	la	verde	Total
green	1/3	1/3	1/3	1
house	1/3	1/3	1/3	1
the	1/3	1/3	1/3	1

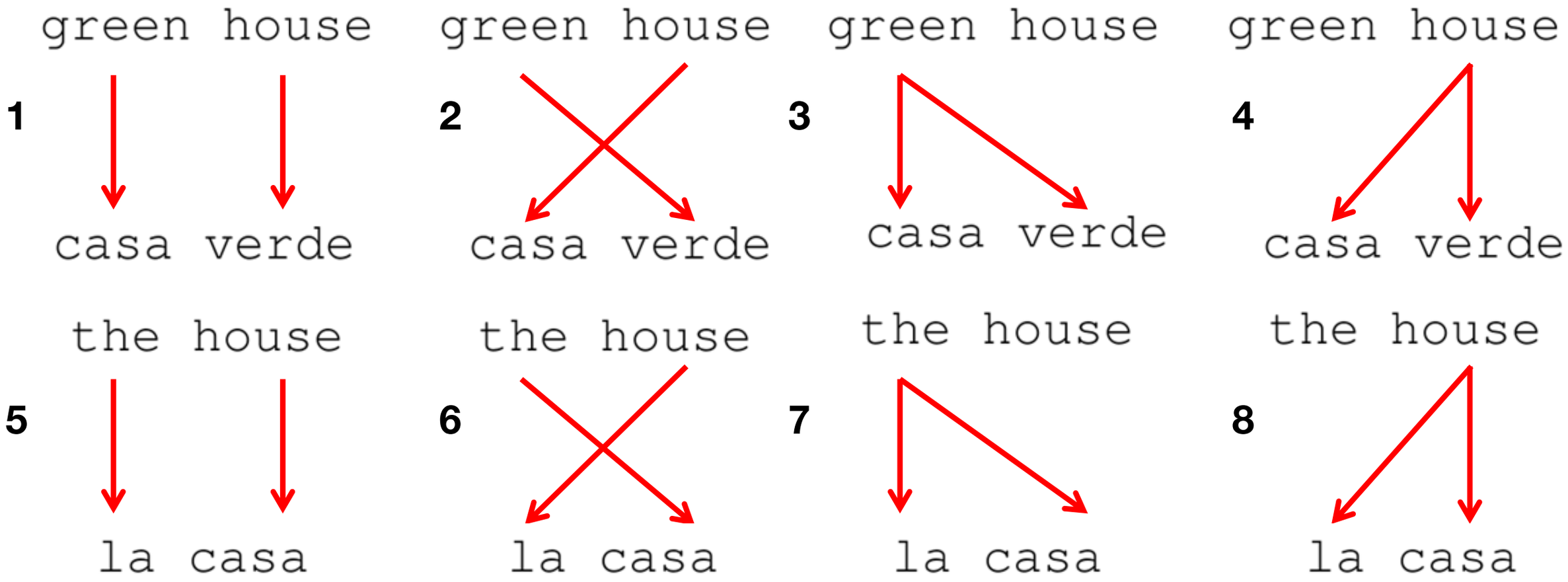
P(a1)	1/9	P(a2)	1/9	P(a3)	1/9	P(a4)	1/9
P(a5)	1/9	P(a6)	1/9	P(a7)	1/9	P(a8)	1/9

$$\hat{P}(F, A|E) = \frac{\epsilon}{(I+1)^J} t(\texttt{casa}|\texttt{green}) t(\texttt{verde}|\texttt{house})$$
$$= \frac{\epsilon}{(2+1)^2} (\frac{1}{3})(\frac{1}{3}) = \frac{\epsilon}{9}\frac{1}{9}$$

(ignoring the ϵ term):

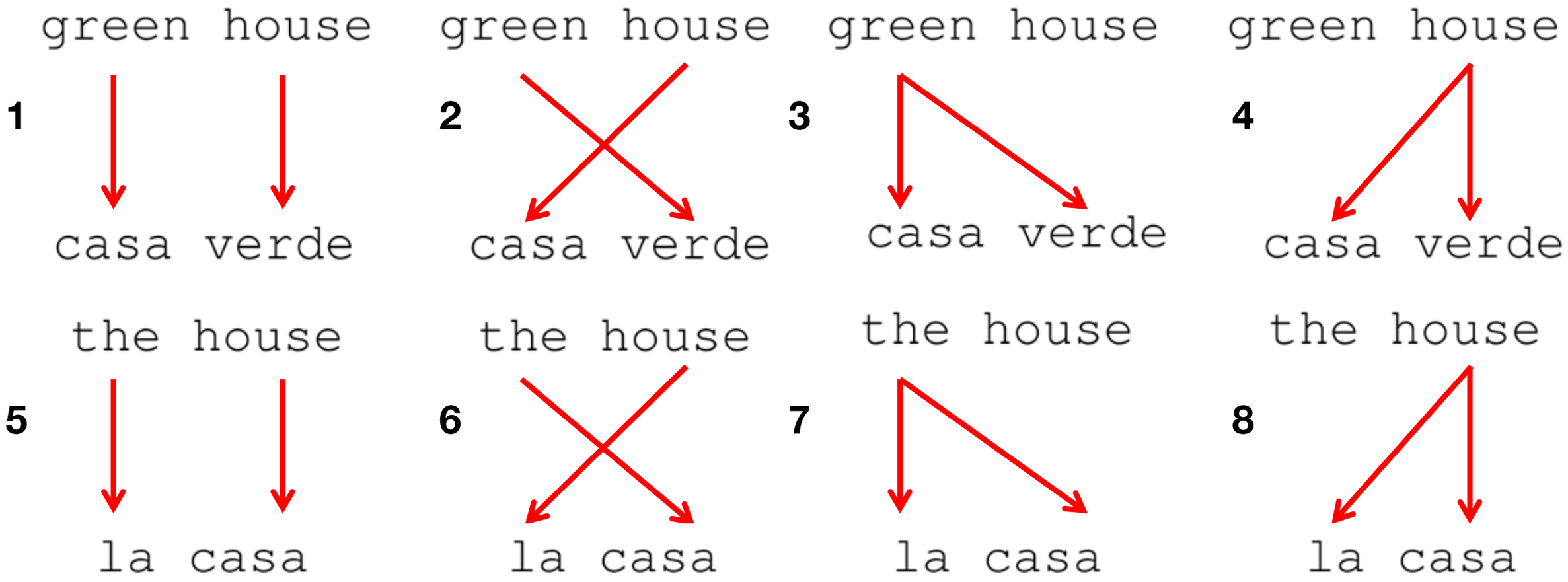
t(B|A)	casa	la	verde	Total
green	1/9 * 2	0	1/9 * 2	4/9
house	1/9 * 4	1/9 * 2	1/9 * 2	8/9
the	1/9 * 2	1/9 * 2	0	4/9

t(B|A)	casa	la	verde	Total
green	1/2	0	1/2	1
house	1/2	1/4	1/4	1
the	1/2	1/2	0	1



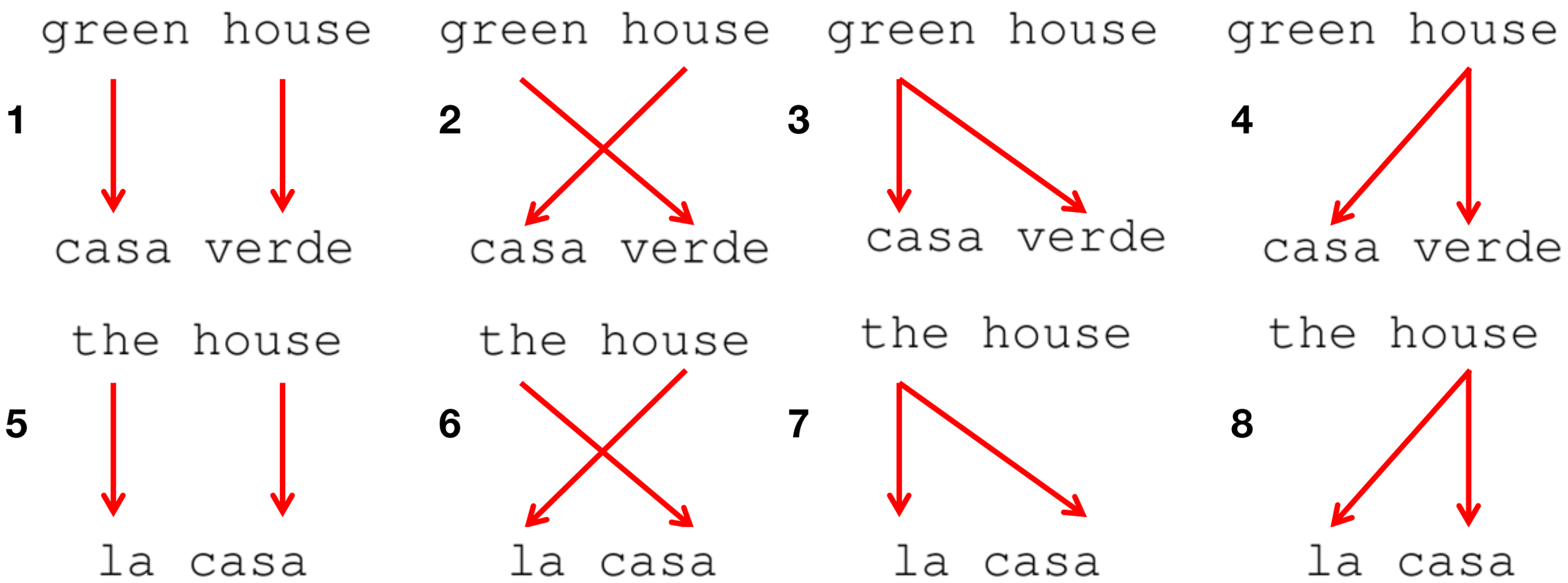
t(B|A)	casa	la	verde	Total
green	1/2	0	1/2	1
house	1/2	1/4	1/4	1
the	1/2	1/2	0	1

P(a1)	P(a2)	P(a3)	P(a4)
P(a5)	P(a6)	P(a7)	P(a8)



t(B|A)	casa	la	verde	Total
green	1/2	0	1/2	1
house	1/2	1/4	1/4	1
the	1/2	1/2	0	1

P(a1)	1/8	P(a2)	1/4	P(a3)	1/4	P(a4)	1/8
P(a5)	1/4	P(a6)	1/8	P(a7)	1/4	P(a8)	1/8



P(a1)	1/8	P(a2)	1/4	P(a3)	1/4	P(a4)	1/8
P(a5)	1/4	P(a6)	1/8	P(a7)	1/4	P(a8)	1/8

t(B|A)	casa	la	verde	Total
green	1/8 + 1/4	0	1/4 + 1/4	7/8
house	1/4 + 1/8 + 1/4 + 1/8	1/8 + 1/8	1/8 + 1/8	10/8
the	1/8 + 1/4	1/4 + 1/4	0	7/8