CA4012

Statistical Machine Translation

Week 7: Phrase-based Translation Model

Lecturer: Dimitar Shterionov

Office: L2.15 Phone: 17006906

E-mail: (dimitar.shterionov@dcu.ie)

Lab Tutor: Meghan Dowling, Pintu Lohar

2nd Semester, 2019-2020 Academic Year

Recap: IBM Models

- What's the main deficiency of IBM Model 1?
- In terms of IBM Model 1, assuming the length of a source-side sentence is l_f, the length of a target-side sentence is l_e, how many possible alignments between them (ignoring the NULL word at the source side)?
- What's the main differences between higher IBM Models and IBM Model 1?

Recap: IBM Models

- What’s the main deficiency of IBM Model 1?
- It is weak at the reordering, because it regards all pos sible reorderings as equally likely.

Recap: IBM Models

- In terms of IBM Model 1, assuming the length of a source-side sentence is l_f, the length of a target-side sentence is l_e, how many possible alignments between them (ignoring the NULL word at the source side)?
- $(l_f)^{le}$

Recap: IBM Models

- What are the main differences between higher IBM Models and IBM Model 1?
  - IBM Model 1: lexical translation;
  - IBM Model 2: adds absolute alignment model;
  - IBM Model 3: adds fertility model;
  - IBM Model 4: adds relative alignment model;
  - IBM Model 5: fixes deficiency.

Exercise in Class

Given the following Chinese-English pairs:

S1	S2
yuan	hen yuan
far	far away

The source side is Chinese, and the target side is English. In this question, the *NULL* token is ignored.

Q1:
Assuming only one-to-one alignment is allowed, please list all the possible word alignments for the two sentence pairs.

Exercise in Class

Given the following Chinese-English pairs:

S1	S2
yuan	hen yuan
far	far away

The source side is Chinese, and the target side is English. In this question, the *NULL* token is ignored.

Q1:
Assuming only one-to-one alignment is allowed, please list all the possible word alignments for the two sentence pairs.

a1: yuan – far
a2: hen – far, yuan – away
a3: hen – away, yuan – far

Q2:

Considering all word alignments as above, state what the following translation probabilities will be after **two** iterations of the **Expectation Maximisation** algorithm and **show all the steps** followed to arrive at these values:

$$t(far|yuan)$$
$$t(way|yuan)$$
$$t(far|hen)$$
$$t(away|hen)$$

Solution 1: Normal IBM 1

- Step 0: Initialisation
- Step 1: Expectation – Alignment probability
  - Translation probability under the alignment:
  - Alignment probability for each alignment

Solution 1: Normal IBM 1

- Step 0: Initialisation
- Step 1: Expectation – Alignment probability
  - Translation probability under the alignment:

$$p(e,a|f) = \frac{\varepsilon}{(l_f+1)^{l_e}} \prod_{j=1}^{l_e} t(e_j|f_{a(j)})$$

  - Alignment probability for each alignment

$$p(a|\boldsymbol{e},\boldsymbol{f}) = \frac{p(\boldsymbol{e},a|\boldsymbol{f})}{p(\boldsymbol{e}|\boldsymbol{f})} = \frac{p(e,a|f)}{\sum_a p(\boldsymbol{e},a|\boldsymbol{f})}$$

Step 2: Maximisation

- Collecting fractional counts:

- Estimate new lexical translation probability:

- Iterate until convergence

Step 2: Maximisation

– Collecting factional counts:

$$c(e|f;\boldsymbol{e},\boldsymbol{f}) = \sum_{a} p(a|\boldsymbol{e},\boldsymbol{f}) \sum_{j=1}^{l_e} \delta(e,e_j)\delta(f,f_{a(j)})$$

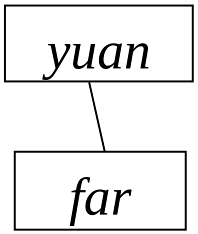
– Estimate new lexical translation probability:

$$t(e|f;\boldsymbol{e},\boldsymbol{f}) = \frac{\sum_{(\boldsymbol{e},\boldsymbol{f})} c(e|f;\boldsymbol{e},\boldsymbol{f})}{\sum_{e}\sum_{(\boldsymbol{e},\boldsymbol{f})} c(e|f;\boldsymbol{e},\boldsymbol{f})}$$

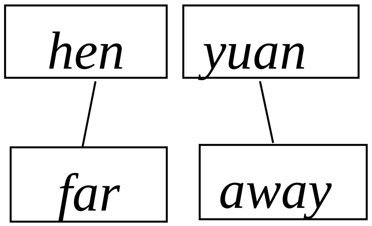
- Iterate until convergence

Q1:

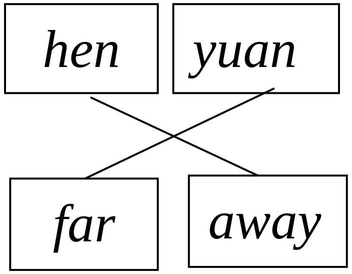
a1:



a2:



a3:



Q2

- Initialisation:
  - Input words W1={hen, yuan}
  - Output words W2={far, away}
  - t(far|yuan) = 1/2
  - t(away|yuan) = 1/2
  - t(far|hen) = 1/2
  - t(away|hen) = 1/2

Iteration 1:

- Step 1 - Expectation:
  - Compute the translation probability under each p ossible alignment:
    - p(e, a1|f) = t(far|yuan) = 1/2
    - p(e, a2|f) = t(far|hen)*t(away|yuan) = 1/2 * (1/2) = 1/4
    - p(e, a3|f) = t(away|hen)*t(far|yuan) = 1/2 * (1/2) = 1/4

Iteration 1:

- Step 1 - Expectation:
  - Normalize the alignment probability:
    - p(a1|e,f) = (1/2)/(1/2) = 1
    - p(a2|e,f) = (1/4)/(1/4*2) = 1/2
    - p(a3|e,f) = (1/4)/(1/4*2) = 1/2

Iteration 1:

- Step 2 - Maximisation:
  - Collect counts
    - c(far|yuan) = 1*1+1/2*1 = 3/2
    - c(away|yuan) = 1/2 *1 = 1/2
    - c(far|hen) = 1/2*1= 1/2
    - c(away|hen) = 1/2*1= 1/2

Iteration 1:

- Step 2 - Maximisation:
  - Normalize and estimate model parameters
    - t(far|yuan) = 3/2/(3/2+1/2) = 3/4
    - t(away|yuan) = 1/2/(3/2+1/2) = 1/4
    - t(far|hen) = 1/2/(1/2+1/2) = 1/2
    - t(away|hen) = 1/2/(1/2+1/2) = 1/2

Iteration 2:

- Step 1 - Expectation:
  - Compute the translation probability under one alignment:
    - p(e, a1|f) = t(far|yuan) = 3/4
    - p(e, a2|f) = t(far|hen)*t(away|yuan) = 1/2 * (1/4) = 1/8
    - p(e, a3|f) = t(away|hen)*t(far|yuan) = 1/2 * (3/4) = 3/8

Iteration 2:

- Step 1 - Expectation:
  - Compute the translation probability under one ali gnment:
    - p(e, a1|f) = t(far|yuan) = 3/4
    - p(e, a2|f) = t(far|hen)*t(away|yuan) = 1/2 * (1/4) = 1/8
    - p(e, a3|f) = t(away|hen)*t(far|yuan) = 1/2 * (3/4) = 3/8

Iteration 2:

- Step 1 - Expectation:
  - Normalize the alignment probability:
    - p(a11|e,f) = (3/4)/(3/4) = 1
    - p(a22|e,f) = (1/8)/(4/8) = 1/4
    - p(a23|e,f) = (3/8)/(4/8) = 3/4

Iteration 2:

- Step 2 - Maximisation:
  - Collect counts
    - c(far|yuan) = 1*1+3/4*1 = 7/4
    - c(away|yuan) = 1/4*1 = 1/4
    - c(far|hen) = 1/4*1+ = 1/4
    - c(away|hen) = 3/4*1=3/4

Iteration 2:

- Step 2 - Maximisation:
  - Normalize and estimate model parameters
    - t(far|yuan) = 7/4/(7/4+1/4)=7/8
    - t(away|yuan) = 1/4/(8/4)=1/8
    - t(far|hen) = 1/4/(1/4+3/4)=1/4
    - t(away|hen) = 3/4/(1/4+3/4)=3/4

Results after two Iterations

- t(far|yuan) = 7/4/(7/4+1/4)=7/8
- t(away|yuan) = 1/4/(8/4)=1/8
- t(far|hen) = 1/4/(1/4+3/4)=1/4
- t(away|hen) = 3/4/(1/4+3/4)=3/4

Solution 2: Simplified IBM 1

Step 1: Initialize model parameters $p(e|f)$

Step 2: Collect counts for word pair (e, f)

Step 3: Estimate new model parameters

Iterate untill convergence.

Solution 2: Simplified IBM 1

Step 1: Initialize model parameters $p(e|f)$

Step 2: Collect counts for word pair (e, f)

$$c(e|f;\boldsymbol{e},\boldsymbol{f}) = \frac{t(e|f)}{\sum_{i=0}^{l_f} t(e|f_i)} \sum_{j=1}^{l_e} \delta(e, e_j) \sum_{i=0}^{l_f} \delta(f, f_i)$$

Step 3: Estimate new model parameters

$$t(e|f;\boldsymbol{e},\boldsymbol{f}) = \frac{\sum_{(\boldsymbol{e},\boldsymbol{f})} c(e|f;\boldsymbol{e},\boldsymbol{f})}{\sum_e \sum_{(\boldsymbol{e},\boldsymbol{f})} c(e|f;\boldsymbol{e},\boldsymbol{f})}$$

Iterate untill convergence.

Q2

- Initialisation:
  - Input words W1={hen, yuan}
  - Output words W2={far, away}
    - t(far|yuan) = 1/2
    - t(away|yuan) = 1/2
    - t(far|hen) = 1/2
    - t(away|hen) = 1/2

Iteration 1:

- Collect counts for word pairs sentence by sentence
  - S1: yuan - far
    - c(far|yuan) = $\frac{1/2}{1/2} * 1 = 1$
  - S2: hen yuan – far away
    - c(far|yuan) = $\frac{1/2}{\frac{1}{2}+\frac{1}{2}} * 1 = \frac{1}{2}$
    - c(away|yuan) = $\frac{1/2}{\frac{1}{2}+\frac{1}{2}} * 1 = \frac{1}{2}$
    - c(far|hen) = $\frac{1/2}{\frac{1}{2}+\frac{1}{2}} * 1 = \frac{1}{2}$
    - c(away|hen) = $\frac{1/2}{\frac{1}{2}+\frac{1}{2}} * 1 = \frac{1}{2}$

Iteration 1:

- Using Simplified IBM Model 1 to estimate new word translation probabilities
  - t(far|yuan) = $\frac{1+1/2}{\frac{1}{2}+\frac{1}{2}+1} = \frac{3}{4}$
  - t(away|yuan) = $\frac{1/2}{\frac{1}{2}+\frac{1}{2}+1} = \frac{1}{4}$
  - t(far|hen) = $\frac{1/2}{\frac{1}{2}+\frac{1}{2}} = \frac{1}{2}$
  - t(away|hen) = $\frac{1/2}{\frac{1}{2}+\frac{1}{2}} = \frac{1}{2}$

Iteration 2:

- Collect counts for word pairs sentence by sentence
  - S1: yuan - far
    - c(far|yuan) = $\frac{3/4}{3/4} * 1 = 1$
  - S2: hen yuan – far away
    - c(far|yuan) = $\frac{3/4}{\frac{3}{4}+\frac{1}{2}} * 1 = \frac{3}{5}$
    - c(away|yuan) = $\frac{1/4}{\frac{1}{4}+\frac{1}{2}} * 1 = \frac{1}{3}$
    - c(far|hen) = $\frac{1/2}{\frac{1}{2}+\frac{3}{4}} * 1 = \frac{2}{5}$
    - c(away|hen) = $\frac{1/2}{\frac{1}{2}+\frac{1}{4}} * 1 = \frac{2}{3}$

Iteration 2:

- Using Simplified IBM Model 1 to estimate new word translation probabilities
  - t(far|yuan) = $\frac{1+3/5}{\frac{3}{5}+\frac{1}{3}+1} = \frac{24}{29}$
  - t(away|yuan) = $\frac{1/3}{\frac{3}{5}+\frac{1}{3}+1} = \frac{5}{29}$
  - t(far|hen) = $\frac{2/5}{\frac{2}{5}+\frac{2}{3}} = \frac{3}{8}$
  - t(away|hen) = $\frac{2/3}{\frac{2}{5}+\frac{2}{3}} = \frac{5}{8}$

Results from two solutions

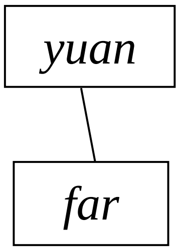
- t(far|yuan) = 7/4/(7/4+1/4)=7/8
- t(away|yuan) = 1/4/(8/4)=1/8
- t(far|hen) = 1/4/(1/4+3/4)=1/4
- t(away|hen) = 3/4/(1/4+3/4)=3/4

- t(far|yuan) = $\frac{1+3/5}{\frac{3}{5}+\frac{1}{3}+1} = \frac{24}{29}$
- t(away|yuan) = $\frac{1/3}{\frac{3}{5}+\frac{1}{3}+1} = \frac{5}{29}$
- t(far|hen) = $\frac{2/5}{\frac{2}{5}+\frac{2}{3}} = \frac{3}{8}$
- t(away|hen) = $\frac{2/3}{\frac{2}{5}+\frac{2}{3}} = \frac{5}{8}$

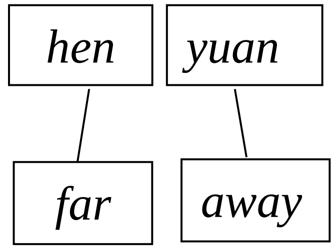
Why different?

Two alignments Missed

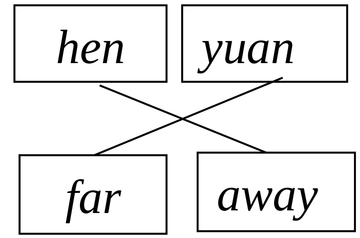
a1:



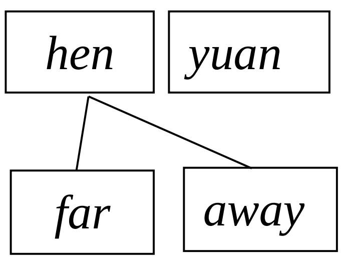
a2:



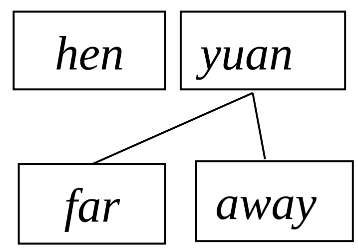
a3:



a4:



a5:



Content

Phrase-based Translation Model

- Learning a Phrase Translation Table
- Bidirectional Word Alignment
- Phrase Pair Extraction
- Phrase Translation Probability

Exercises

Phrase-based Translation Model

- Word-based models translate words as atomic units.

Phrase-based Translation Model

- Word-based models translate words as atomic units.
- Phrase-based models translate phrases as atomic unit s.

Phrase-based Translation Model

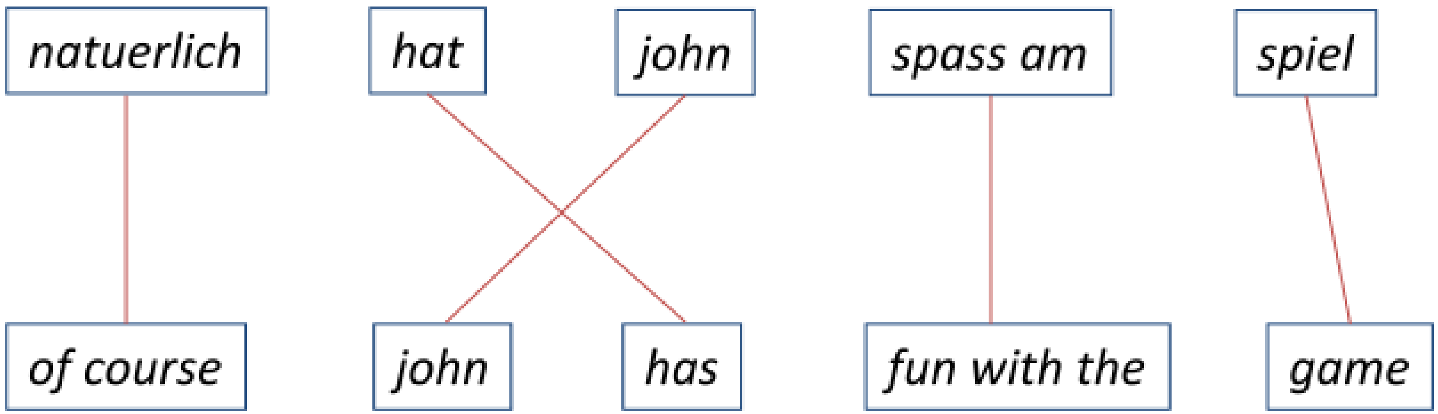
- Word-based models translate words as atomic units.
- Phrase-based models translate phrases as atomic unit s.
- A phrase is a contiguous sequence of words in a sente nce.

Phrase-based Translation Model

- Word-based models translate words as atomic units.
- Phrase-based models translate phrases as atomic unit s.
- A phrase is a contiguous sequence of words in a sente nce.
  - He likes reading

Phrase-based models are the “standard” model in statistical machine translation.
Short: PBSMT, PB-SMT

Example



- Source sentence is segmented into phrases.
- Each phrase is translated into target language.
- Phrases are re-ordered.

Characteristics from Example

- A monolingual phrase:
  - A phrase can be any contiguous sequence of words in a sentence

    e.g. of course, fun with the

Characteristics from Example

- A monolingual phrase:
  - A phrase is not necessarily syntactic well-formed

  e.g. fun with the

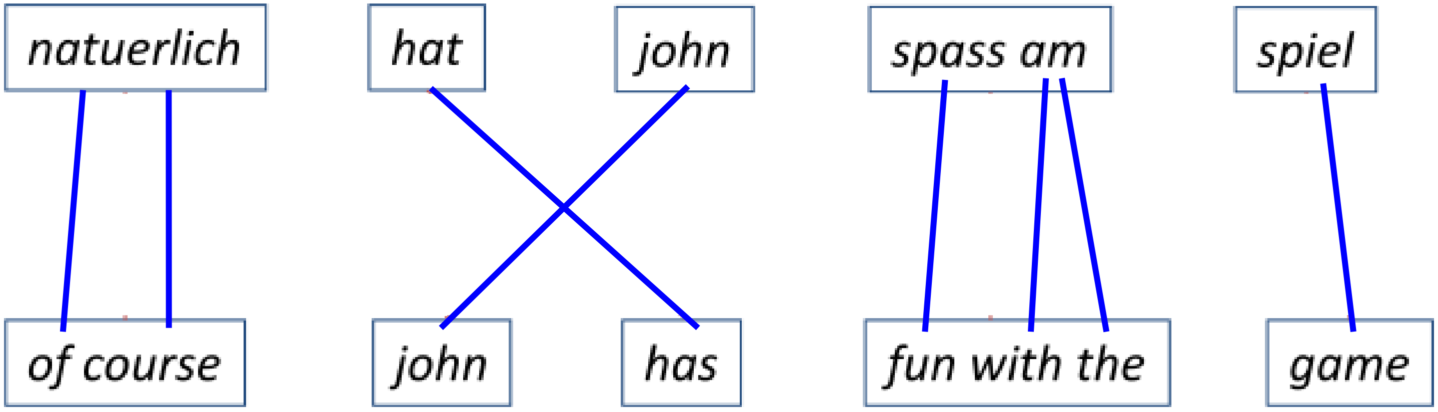
Characteristics from Example

- A monolingual phrase:
  - A phrase is not necessarily semantically meaningful

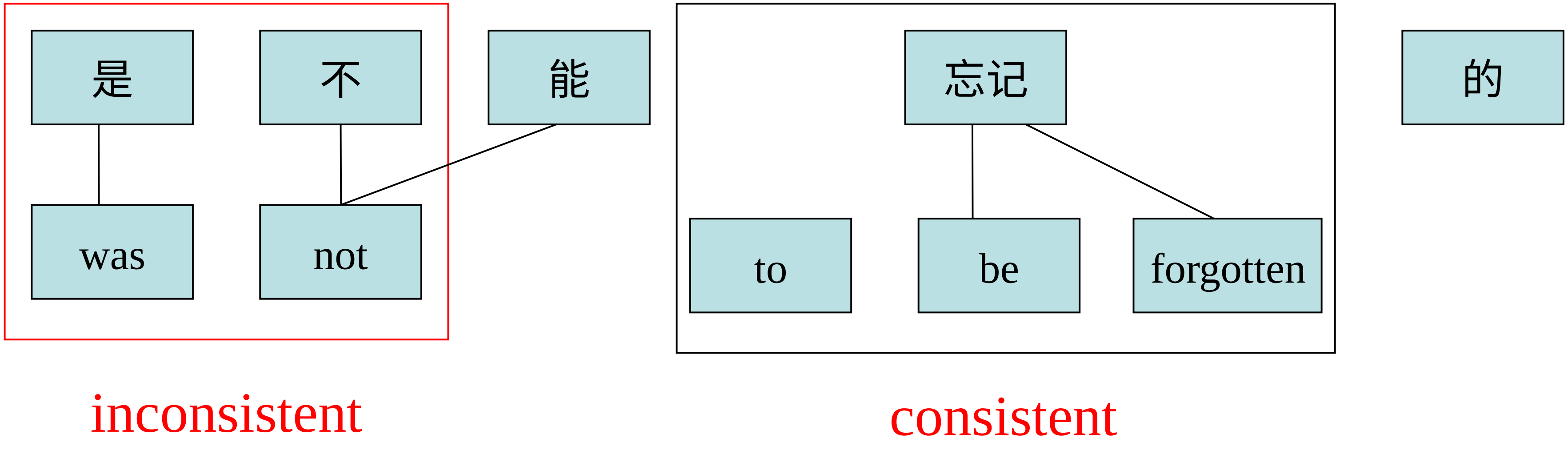
  e.g. with the

Characteristics from Example

- A bilingual phrase pair should be consistent with word alignment.



Bilingual Phrase Pairs: Consistency



Phrase-based Translation Model

- Advantages:

  – many-to-many translation can handle non-compositional phrases or idioms

  e.g. real estate, face value, kick the bucket, shooting the breeze

Phrase-based Translation Model

- Advantages:
  – use of local context in translation

    e.g. the boy in a red shirt | 穿 红衬衣 的 男孩

Phrase-based Translation Model

- Advantages:
  – the more data, the longer phrases can be learned

  e.g. phrase-based SMT is the state-of-the-art

  Nice to meet you, can I have the bill please?

Phrase-based Translation Model

- Advantages:
  - the model is conceptually much simpler.

    e.g. no need the fertility, insertion and deletion in the word-based models.

Phrase Translation Table

- Main knowledge source:
  - table with phrase translations and their probabilities
- Example: phrase translations for natuerlich

Translation	Probability
of course	0.5
naturally	0.3
of course ,	0.15
, of course ,	0.05

Phrase Translation Table

- Real example taken from Europarl for the German phrase den Vorschlag

English	Probability	English	Probability
the proposal	0.6277	the suggestions	0.0114
's proposal	0.1068	the proposed	0.0114
a proposal	0.0341	the motion	0.0091
the idea	0.025	the idea of	0.0091
this proposal	0.0227	the proposal ,	0.0068
proposal	0.0205	its proposal	0.0068
of the proposals	0.0159	it	0.0068
the proposals	0.0159	…….	

Phrase Translation Table

English	Probability	English	Probability
the proposal	0.6277	the suggestions	0.0114
's proposal	0.1068	the proposed	0.0114
a proposal	0.0341	the motion	0.0091
the idea	0.025	the idea of	0.0091
this proposal	0.0227	the proposal ,	0.0068
proposal	0.0205	its proposal	0.0068
of the proposals	0.0159	it	0.0068
the proposals	0.0159	…….	

- ✓ lexical variation (proposal vs suggestions)
- ✓ morphological variation (proposal vs proposals)
- ✓ included function words (the, a, ...)
- ✓ noise (it)

Content

- Phrase-based Translation Model
- **Learning a Phrase Translation Table**
- Bidirectional Word Alignment
- Phrase Pair Extraction
- Phrase Translation Probability
- Exercises

Learning a Phrase Translation Table

Task: learn the model from a parallel corpus

Learning a Phrase Translation Table

Task: learn the model from a parallel corpus

Three stages:

Learning a Phrase Translation Table

Task: learn the model from a parallel corpus

Three stages:

1. word alignment: using IBM models or other method

Learning a Phrase Translation Table

Task: learn the model from a parallel corpus

Three stages:

1. word alignment: using IBM models or other method
2. extraction of phrase pairs

Learning a Phrase Translation Table

Task: learn the model from a parallel corpus

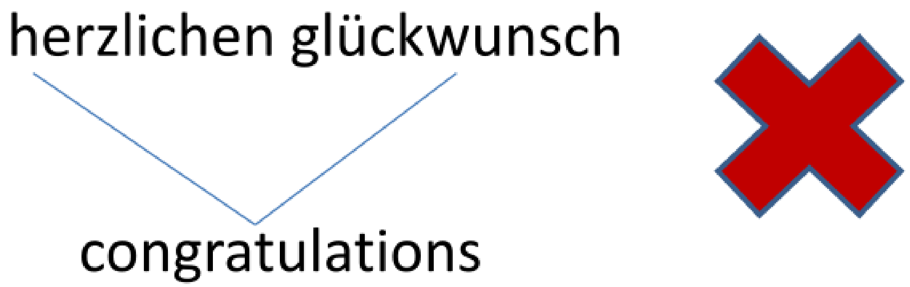
Three stages:

1. word alignment: using IBM models or other method
2. extraction of phrase pairs
3. scoring phrase pairs

Content

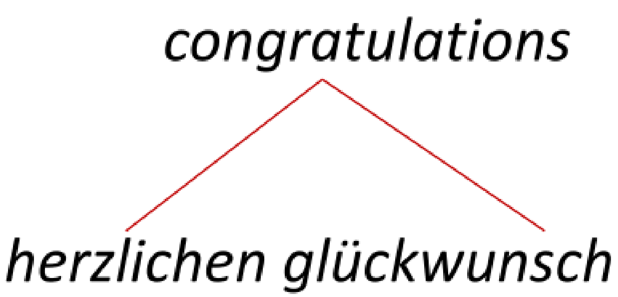
Problems with Word Alignment and IBM Models

- Each target word can be aligned to at most one source word. Therefore, it’s not possible to end up with an alignment of one target word to many source words



How to fix this?

- Compute word alignments in both directions!



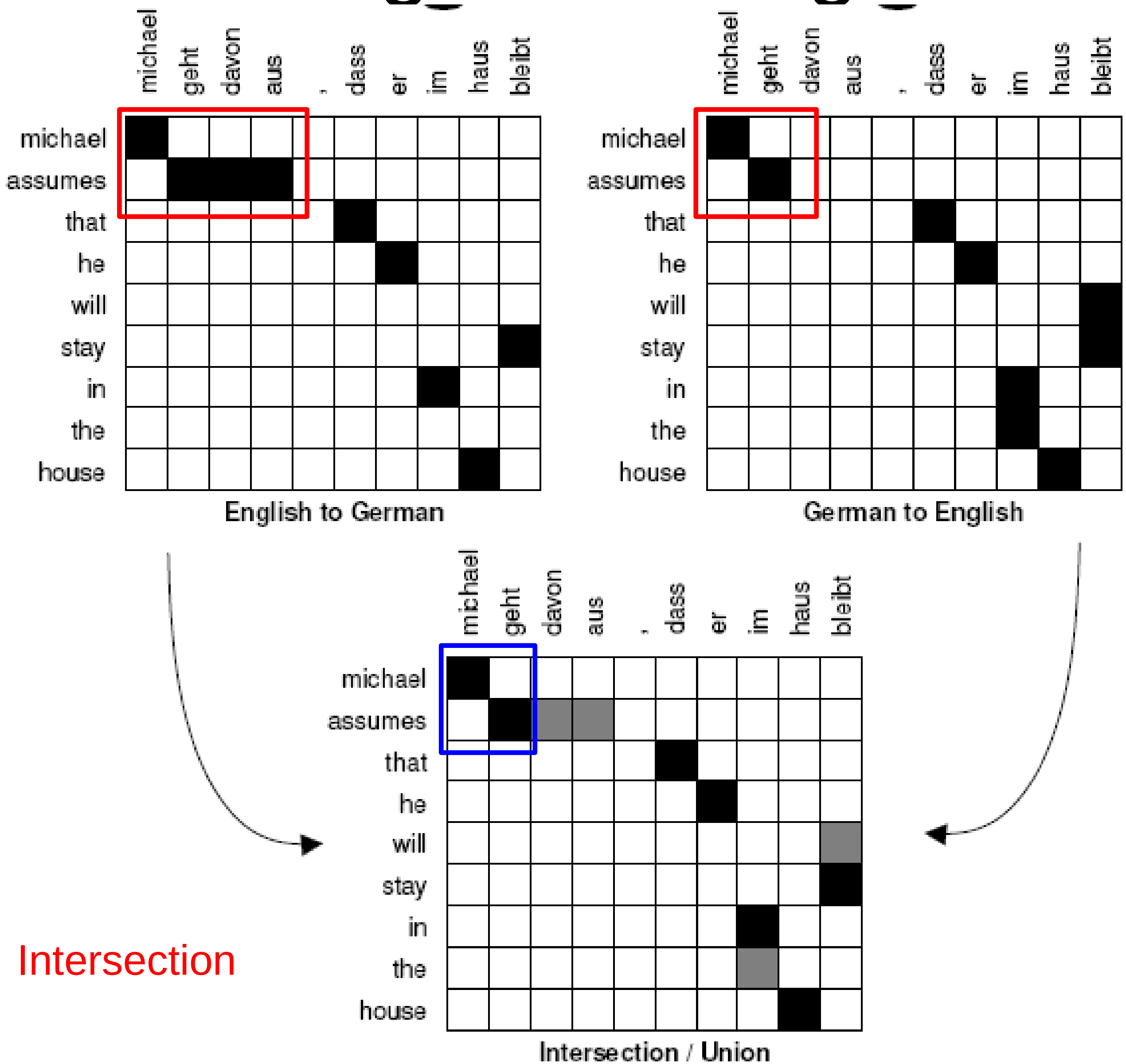
- In this way, we can get many-to-one alignments as well as one-to-many alignments.

Bidirectional Word Alignment

Algorithm of Bidirectional word alignment:

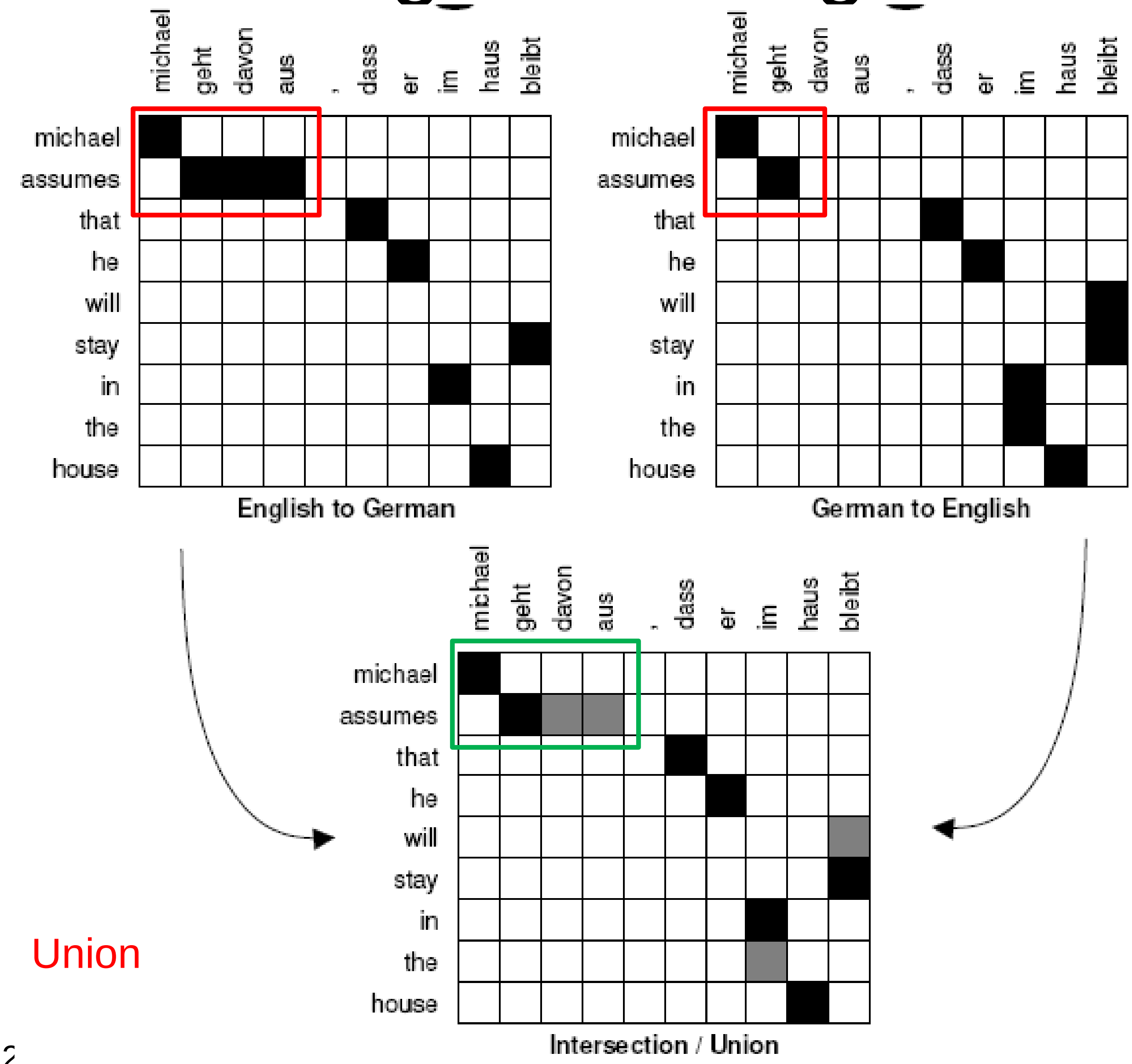
1. Using IBM Models to do word alignment in one direction.
2. Using IBM Models to do word alignment in the other direction.
3. Merge the above two alignments.

Symmetrizing Word Alignments



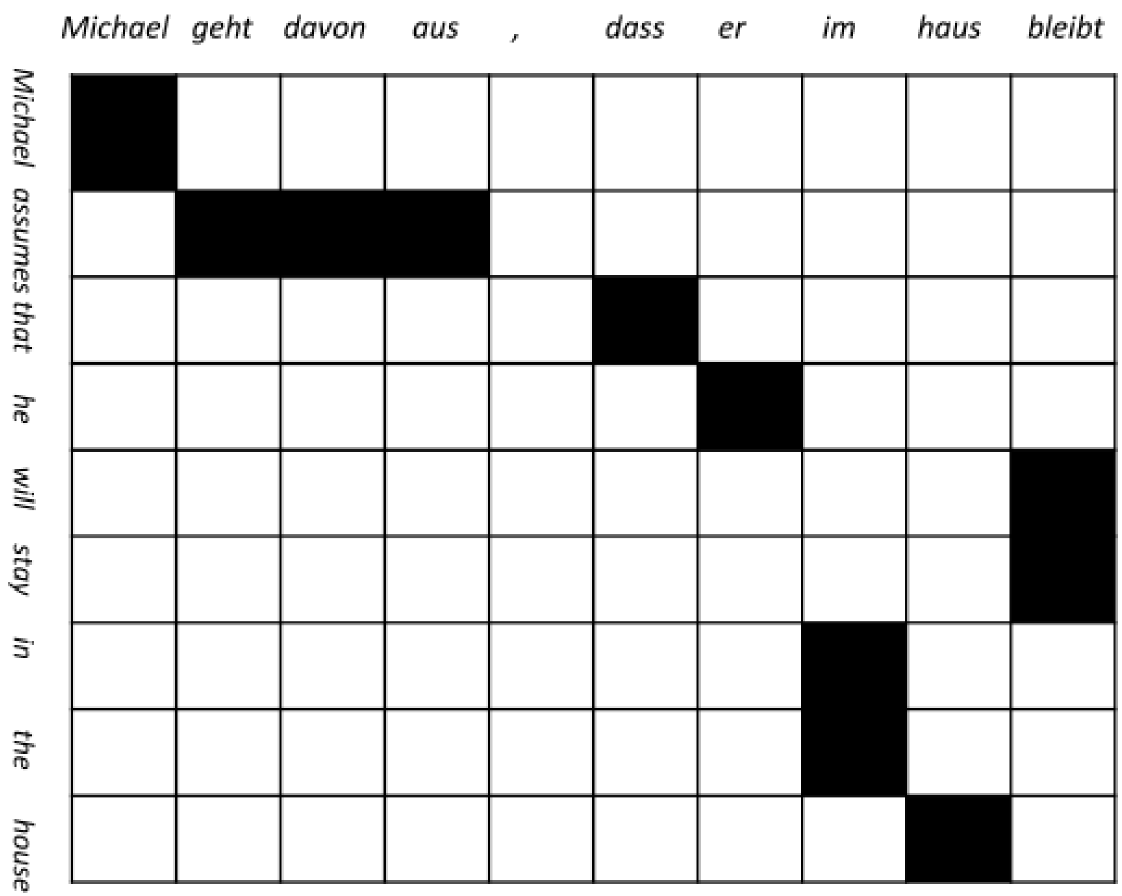
Intersection

Symmetrizing Word Alignments



Union

Bidirectional word alignment



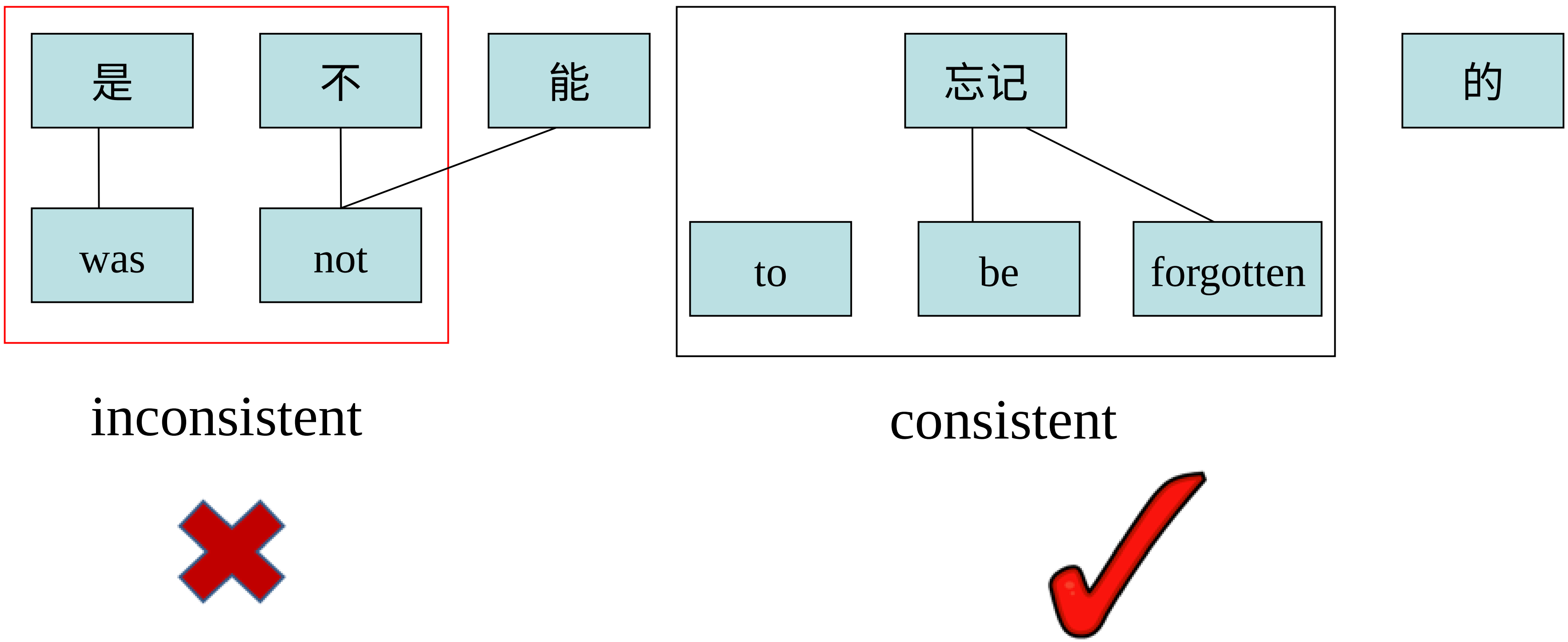
Union

Bidirectional word alignment

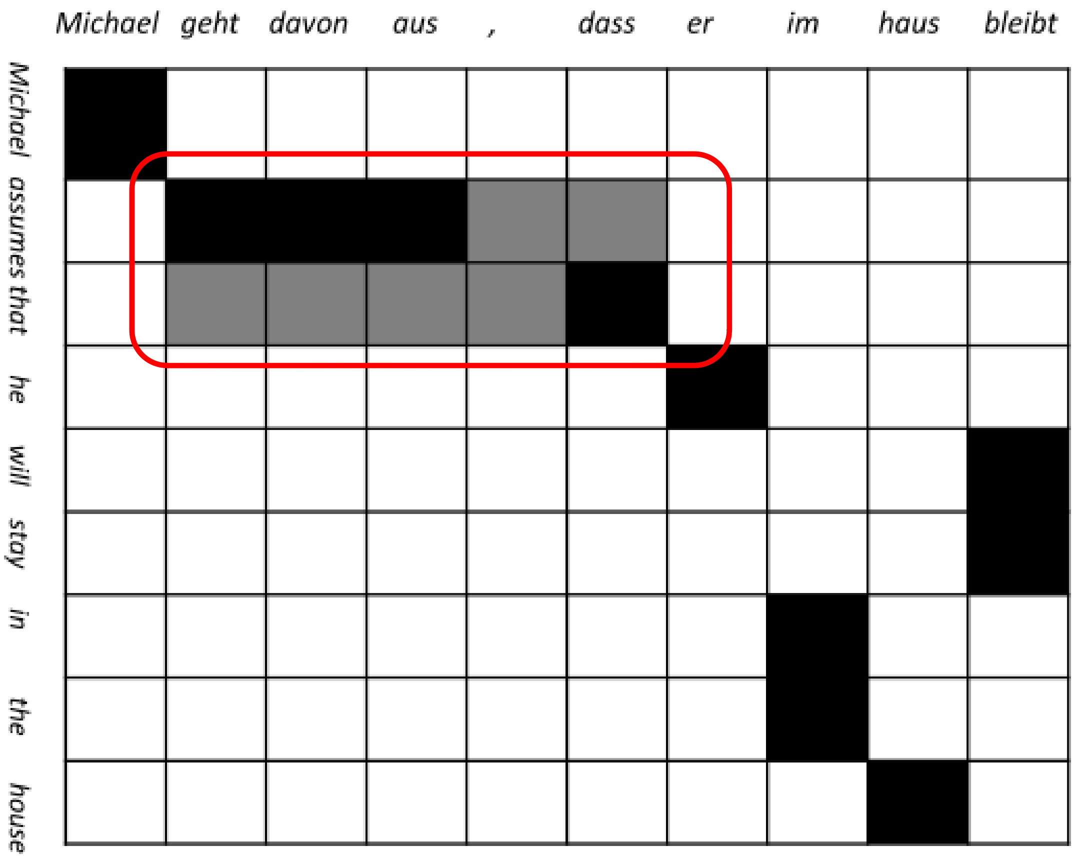
- Add alignment points from union based on heuristics:
  - directly/diagonally neighboring points
  - finally, add alignments that connect unaligned words in source and/or target
- Popular method: grow-diag-final-and

Content

Recall: Bilingual Phrase Pairs



Extracting Phrase Pairs



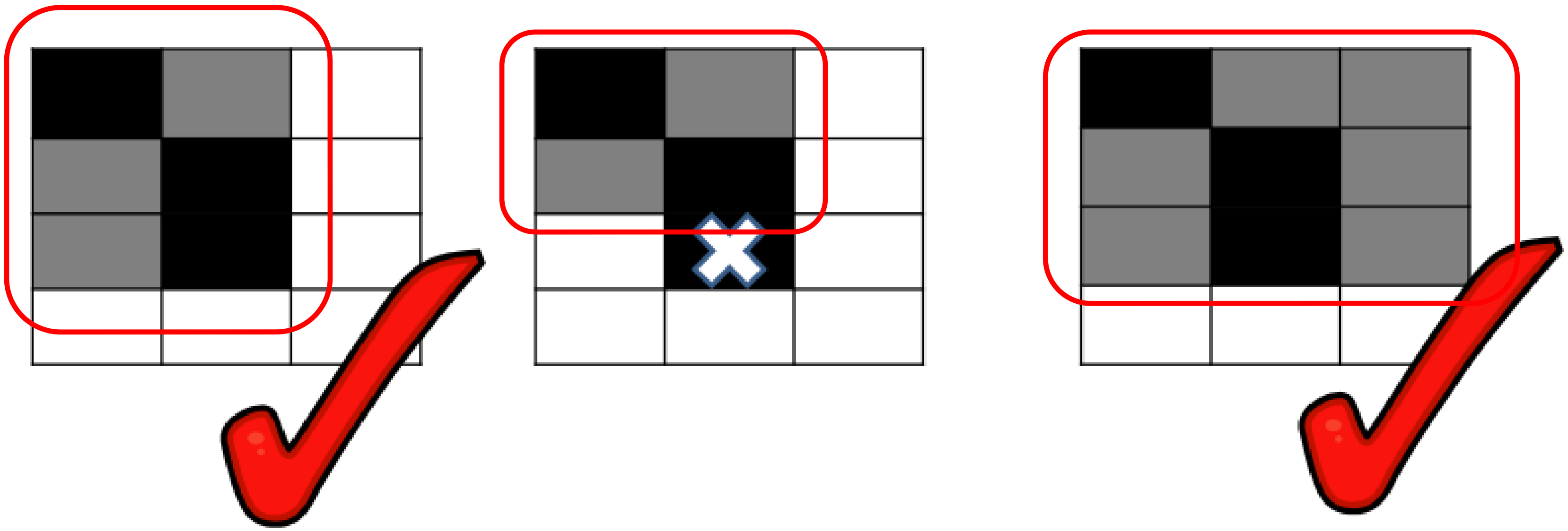
extract phrase pair consistent with word alignment:

assumes that / geht davon aus , dass

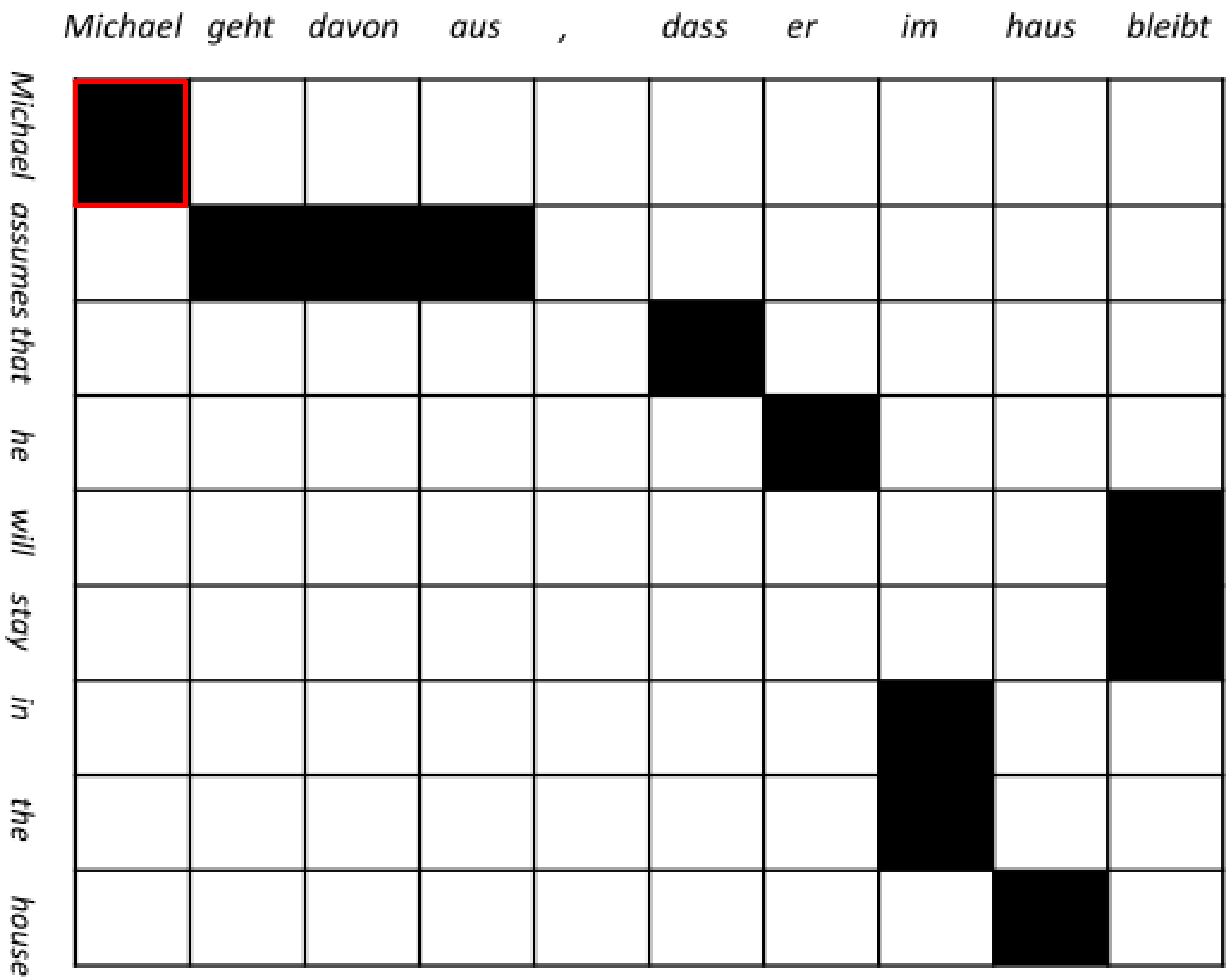
Consistent Conditions

- A phrase pair (e, f) is consistent with a bidirectional word alignment A if and only if
  - For all words e_i in e, if e_i is aligned to a word f_j in A, then f_j is in f.
  - For all words f_j in f, if f_j is aligned to a word e_i in A, then e_i is in e.
  - There exists e_i in e, f_j in f : (e_i, f_j) in A

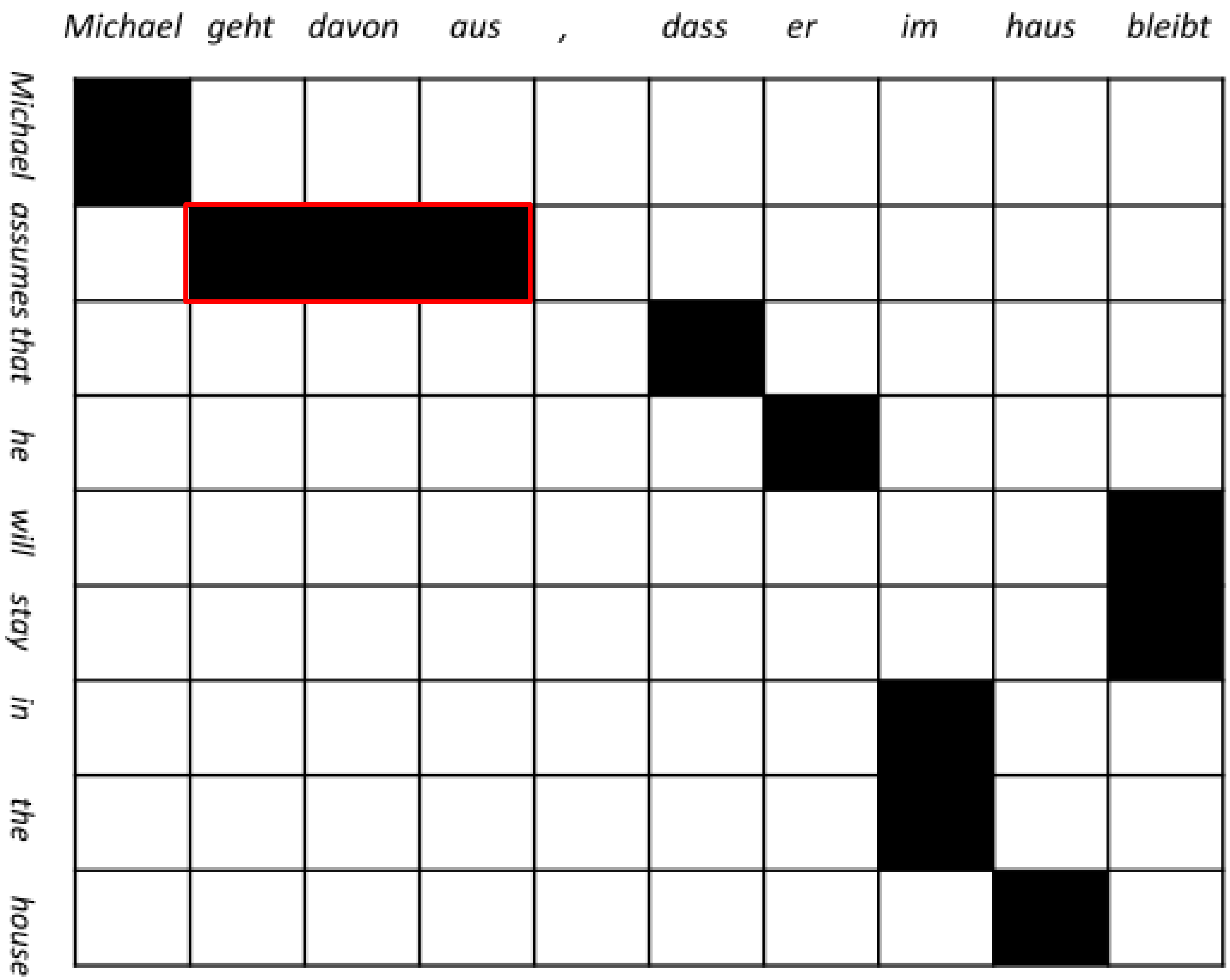
Consistent with Word Alignment



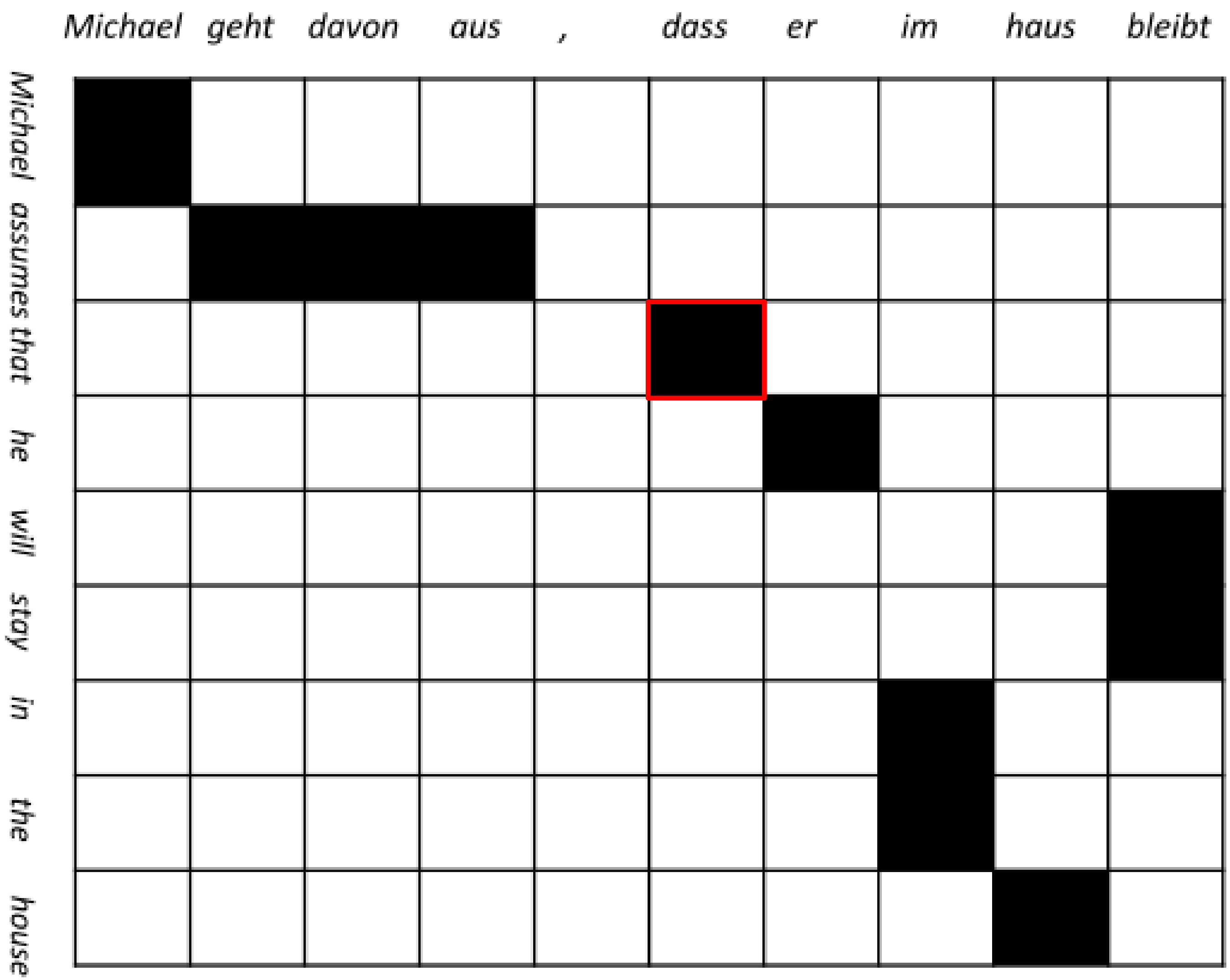
Phrase Pair Extraction



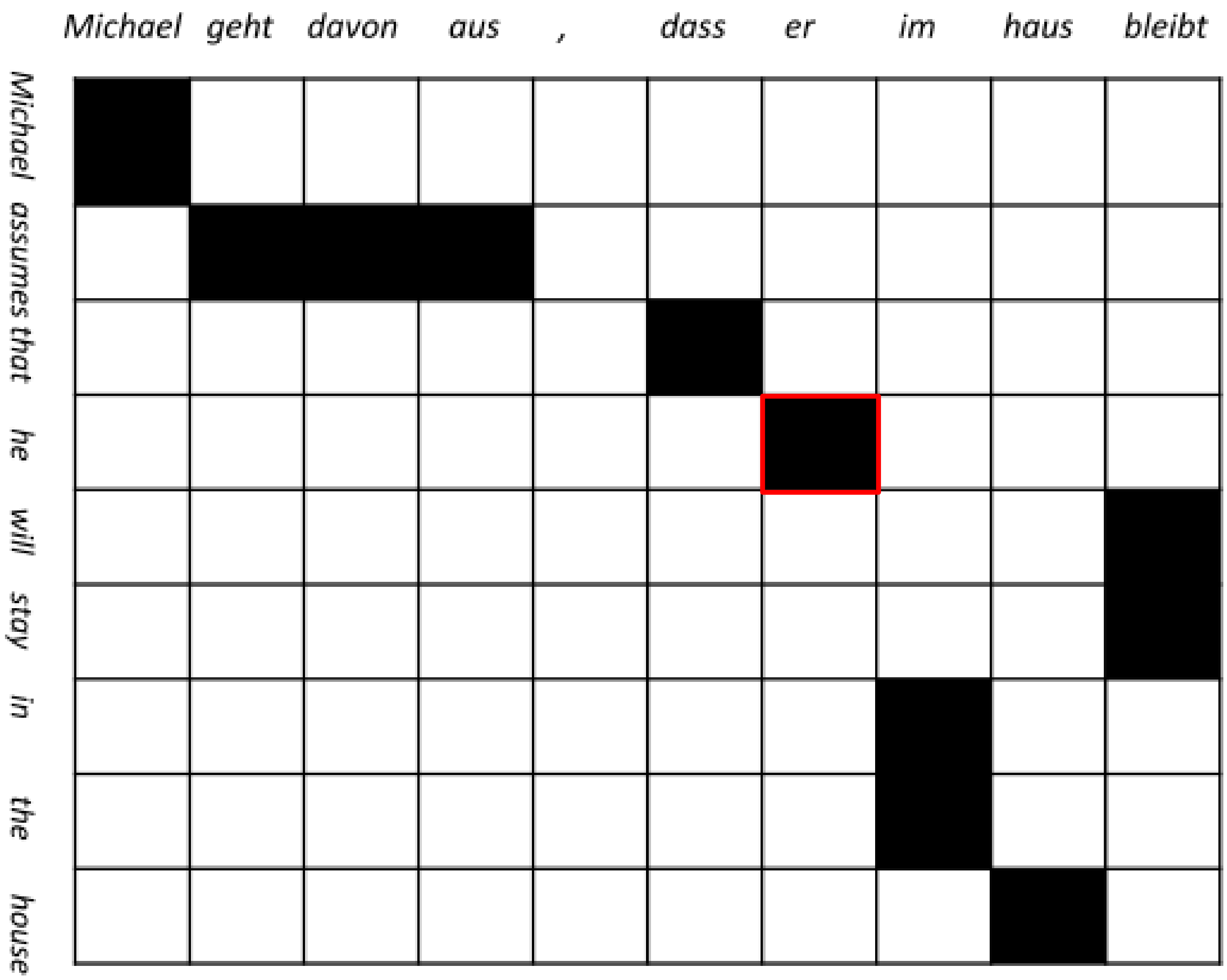
Phrase Pair Extraction



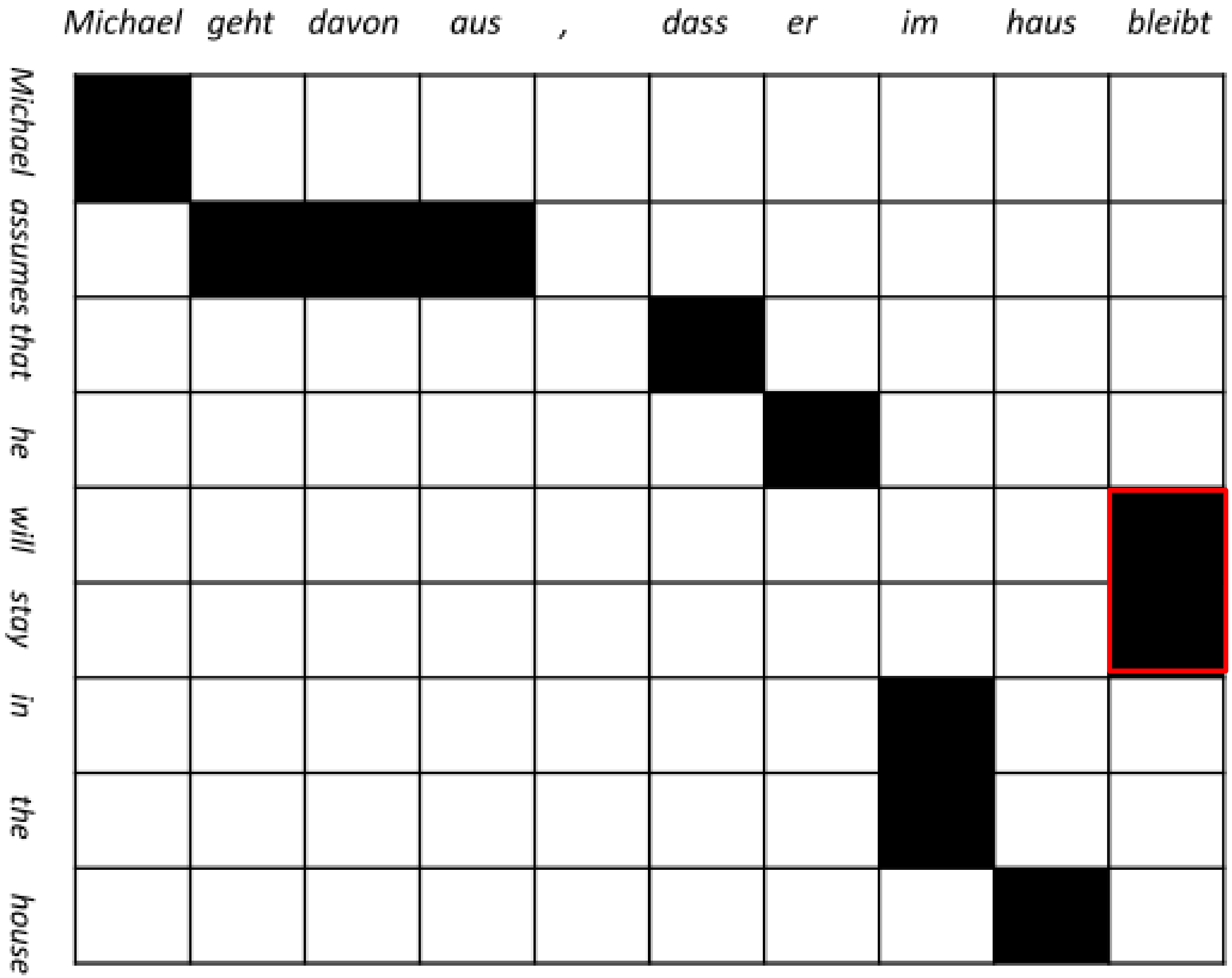
Phrase Pair Extraction



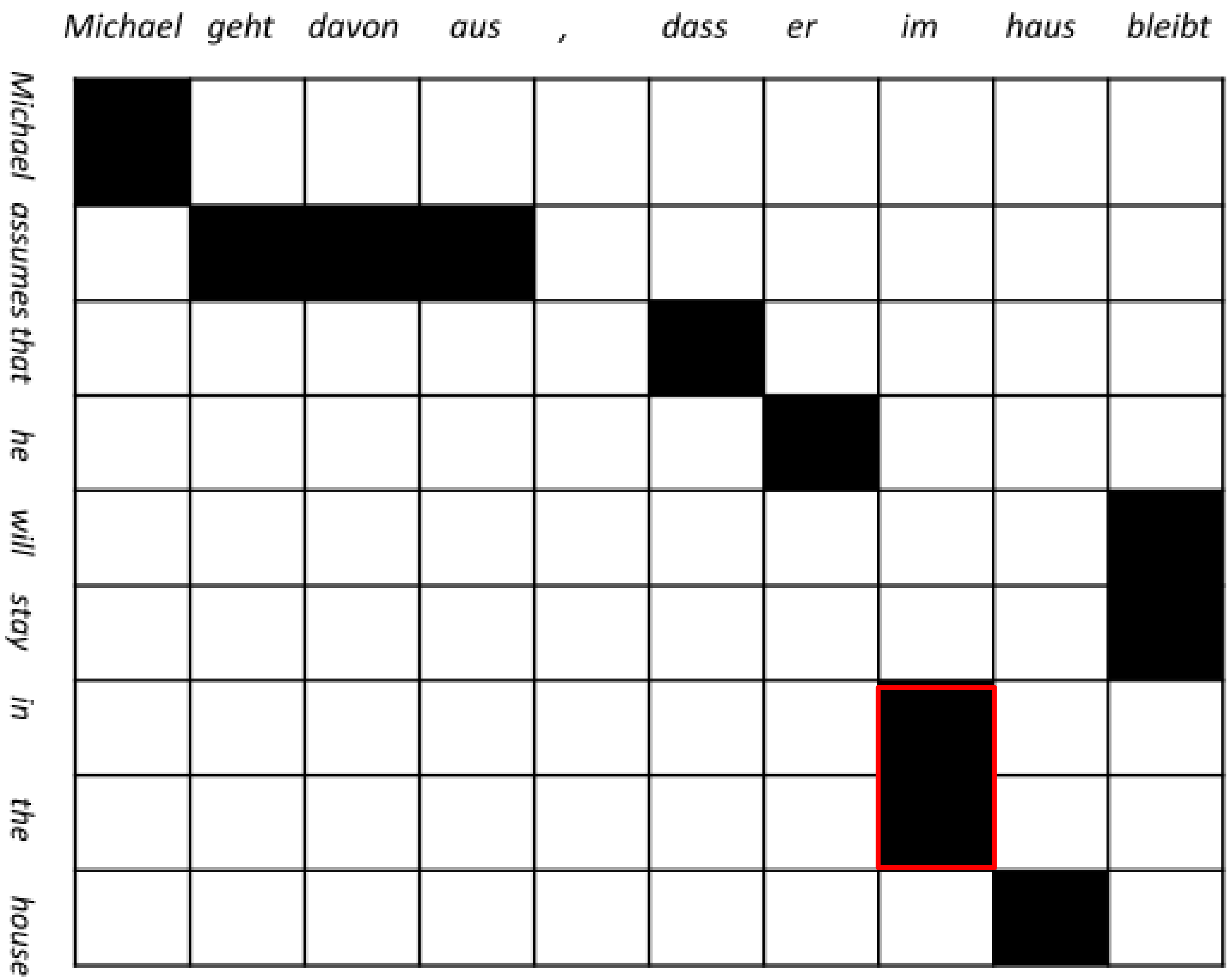
Phrase Pair Extraction



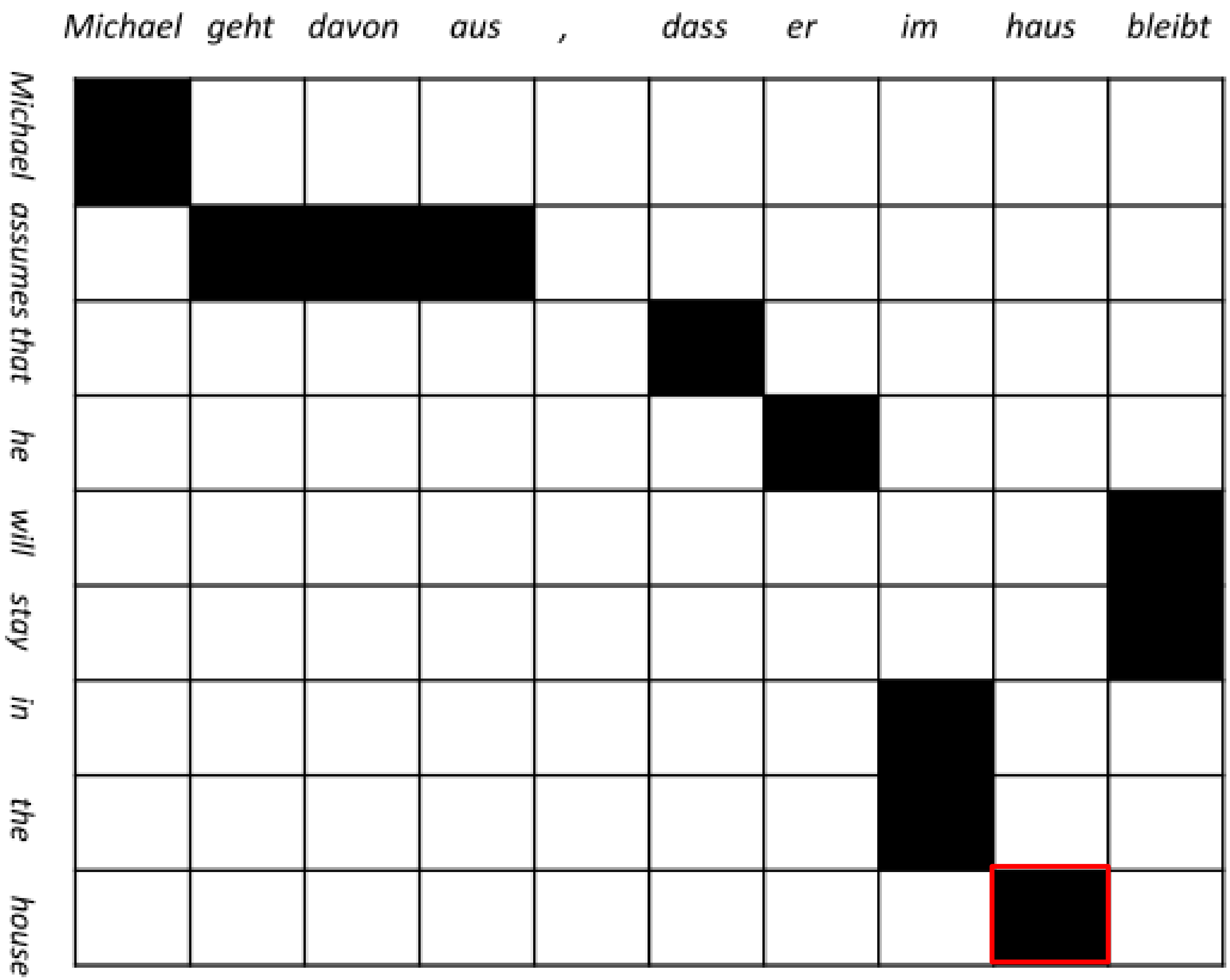
Phrase Pair Extraction



Phrase Pair Extraction



Phrase Pair Extraction



Phrase Pair Extraction

michael | michael

assumes | geht davon aus / geht davon aus ,

that | dass / , dass

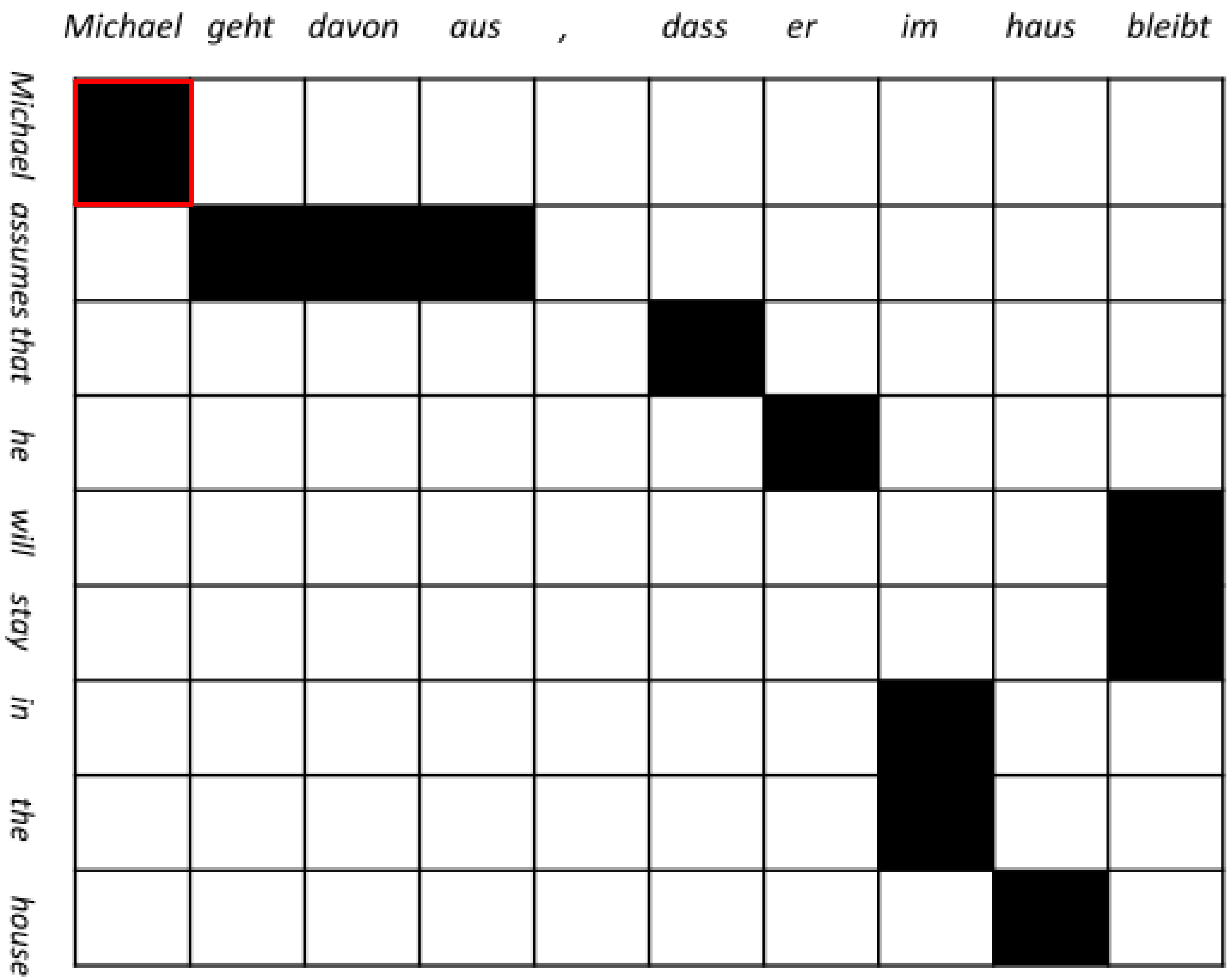
he | er

will stay | bleibt

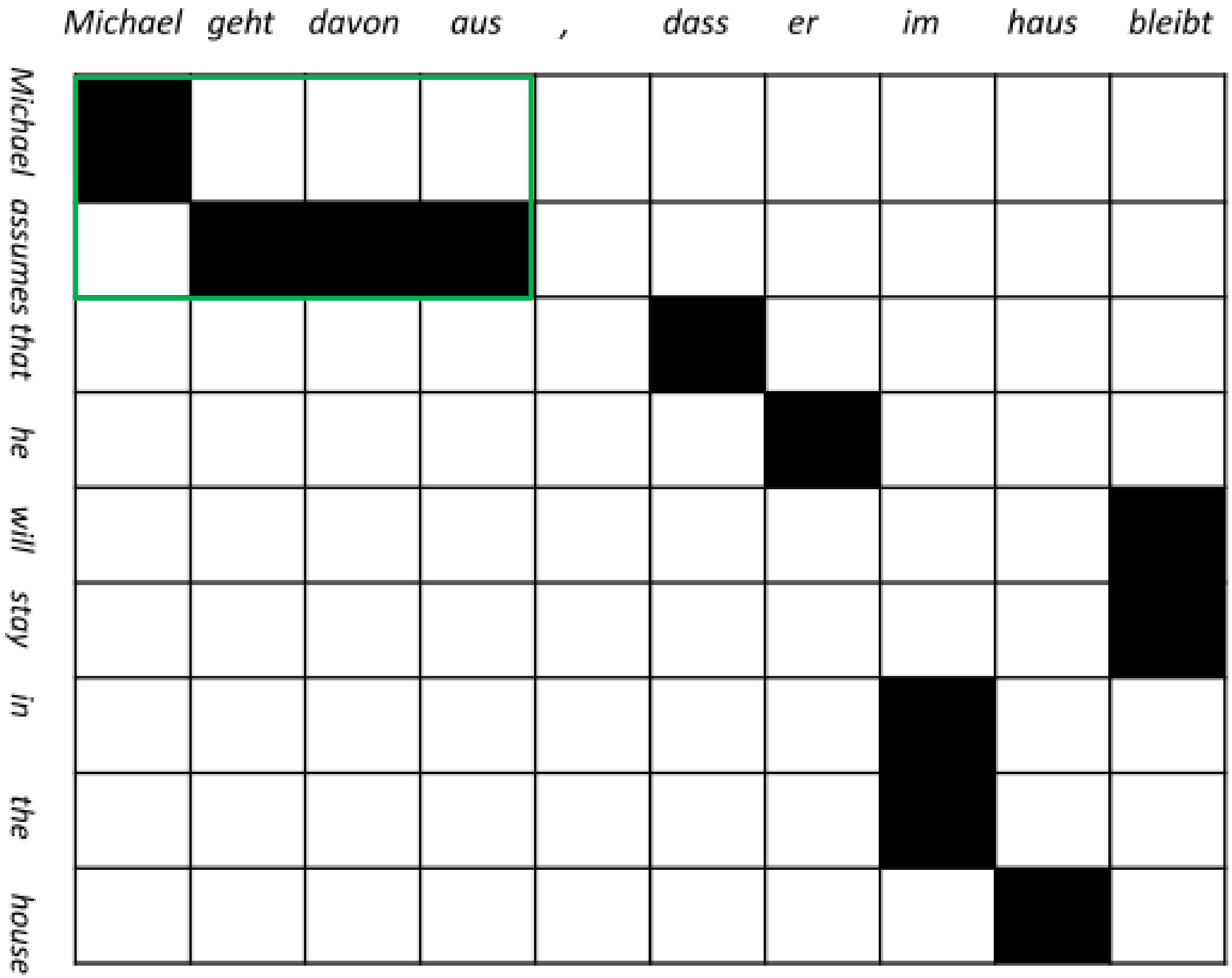
in the | im

house | haus

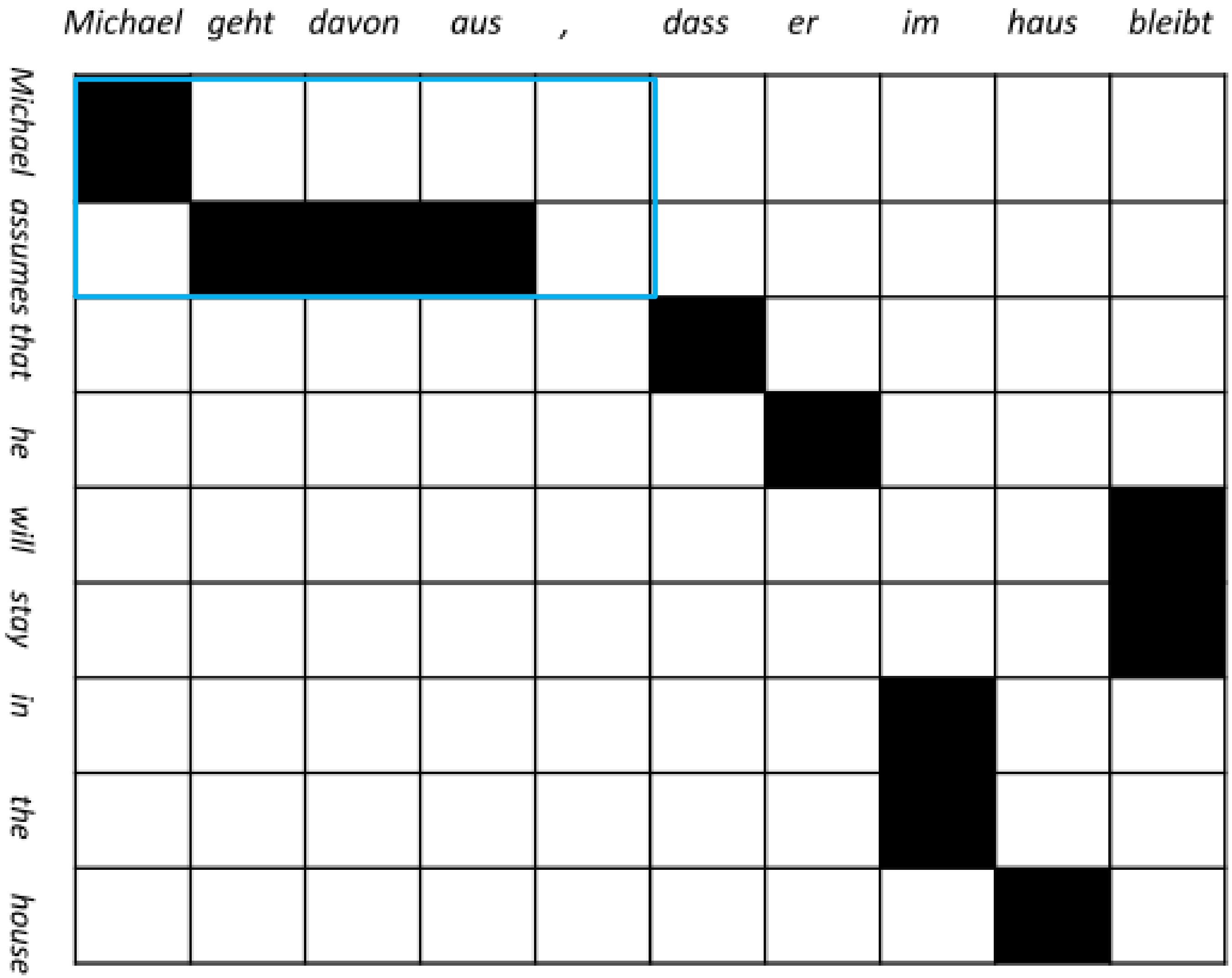
Phrase Pair Extraction



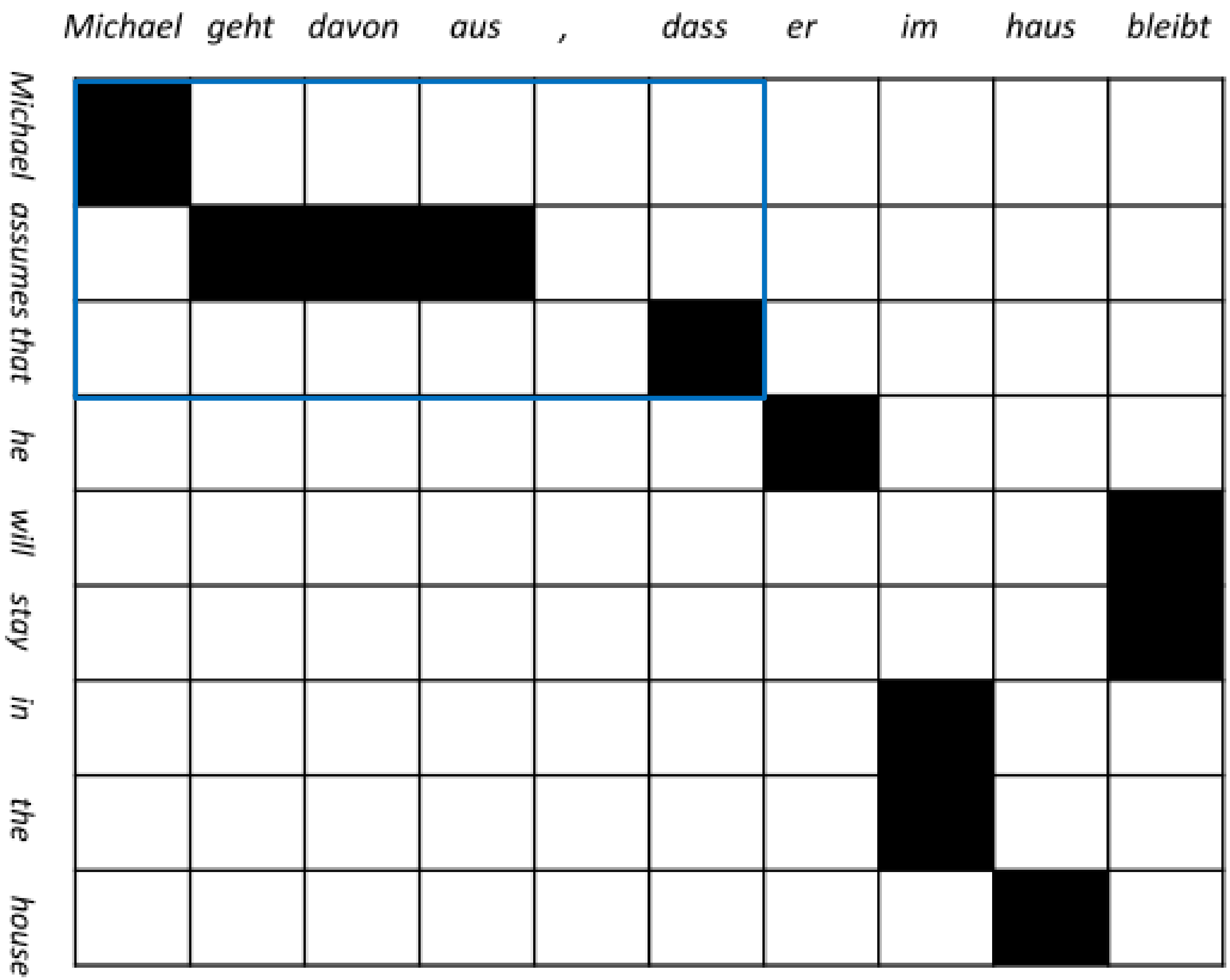
Phrase Pair Extraction



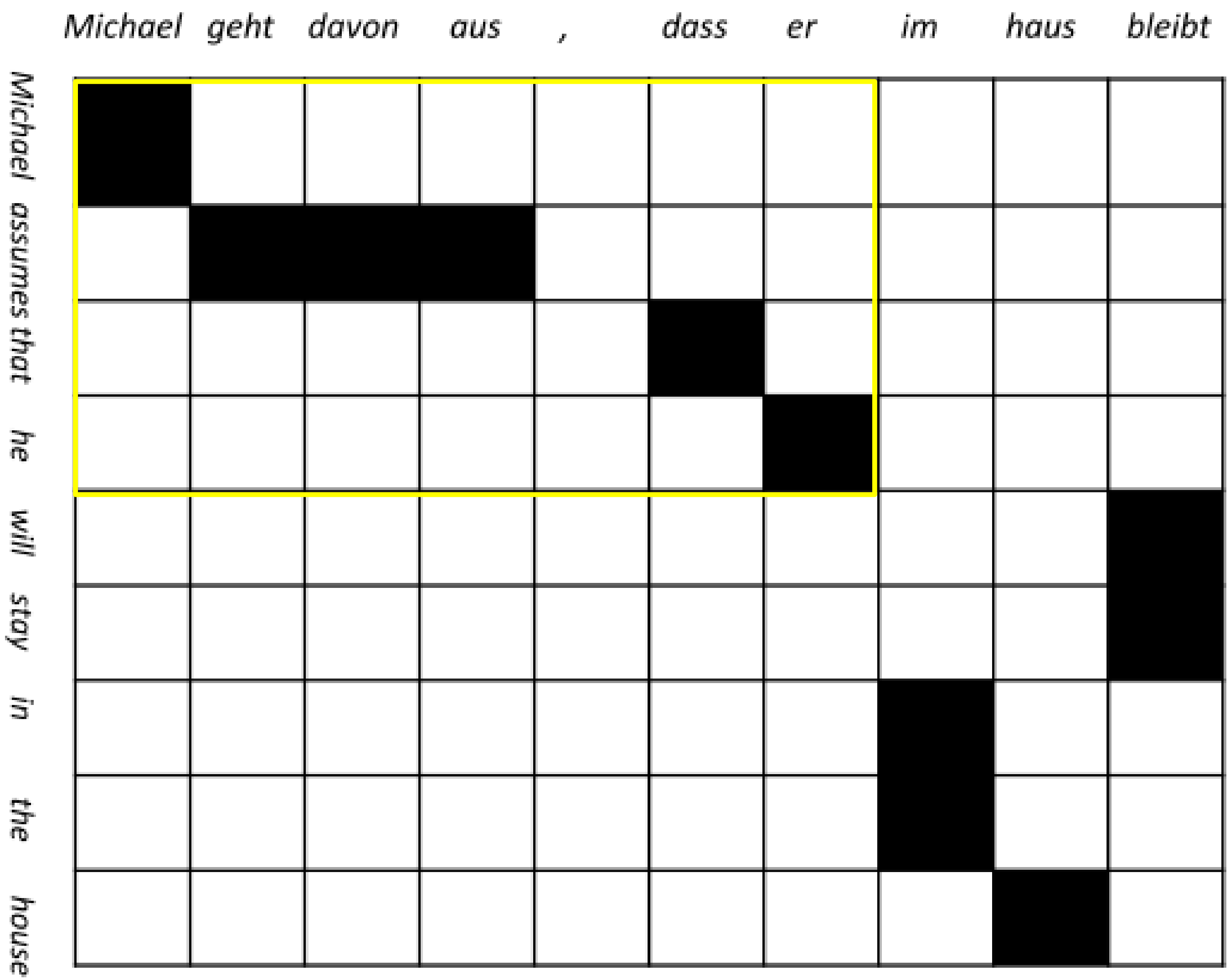
Phrase Pair Extraction



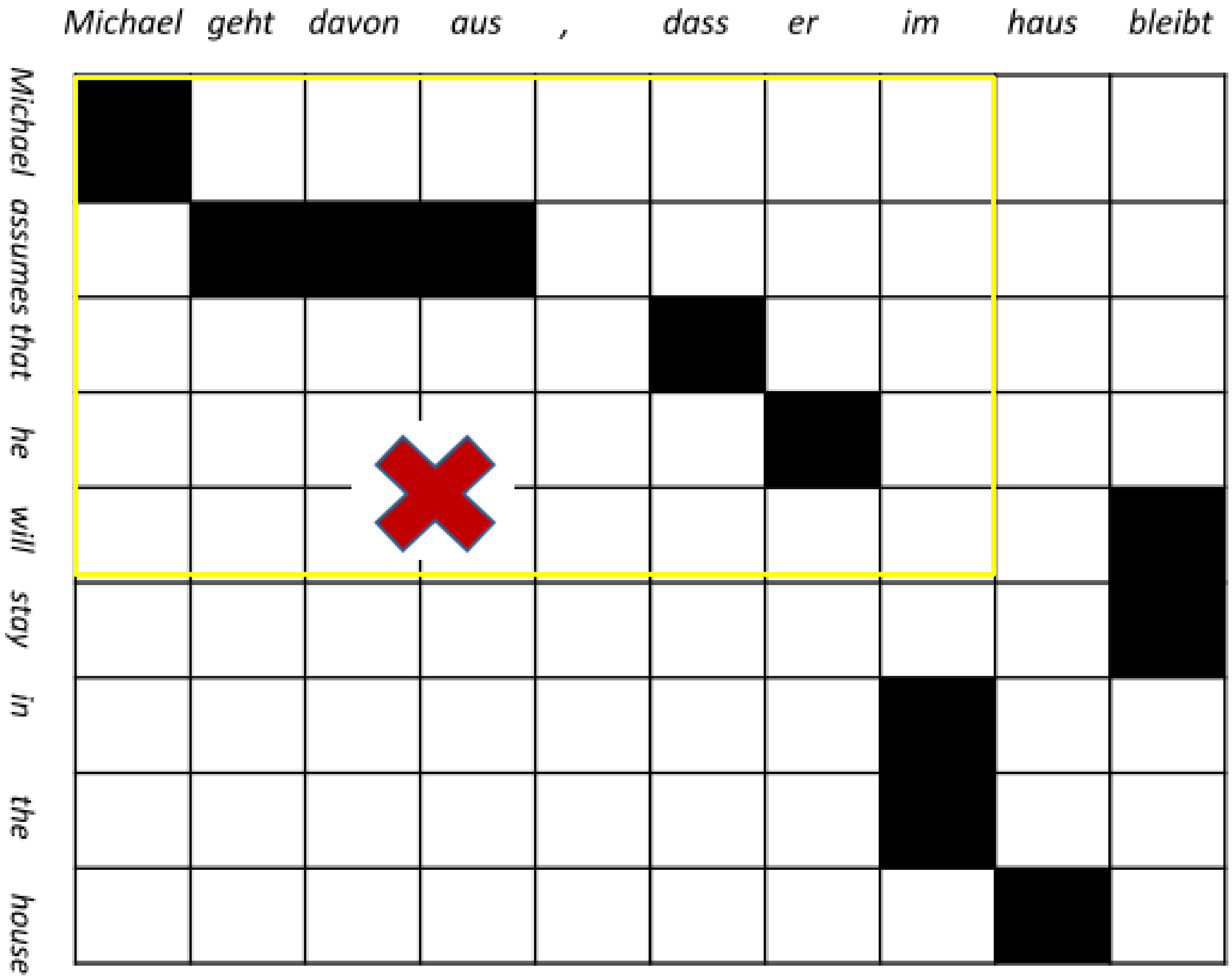
Phrase Pair Extraction



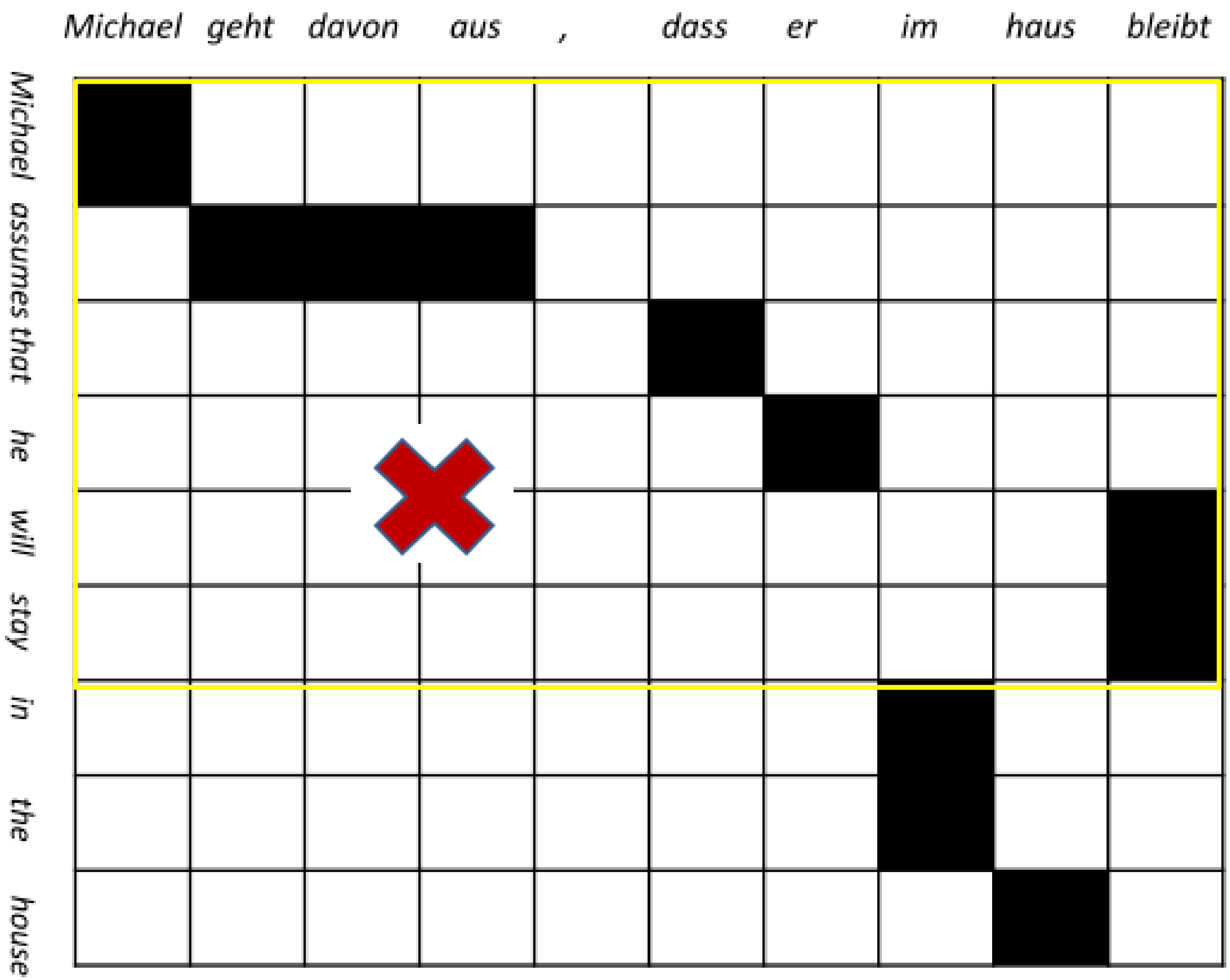
Phrase Pair Extraction



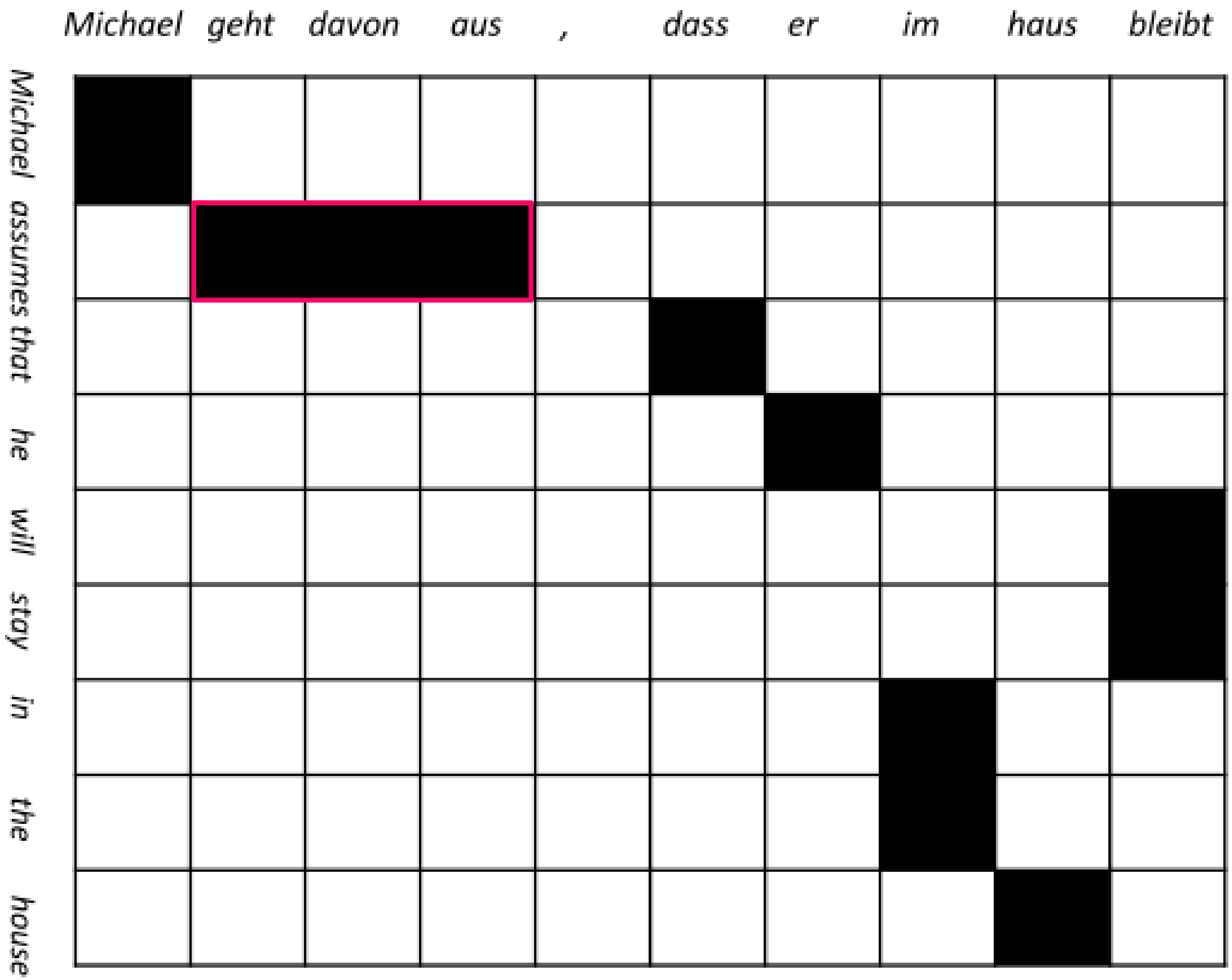
Phrase Pair Extraction



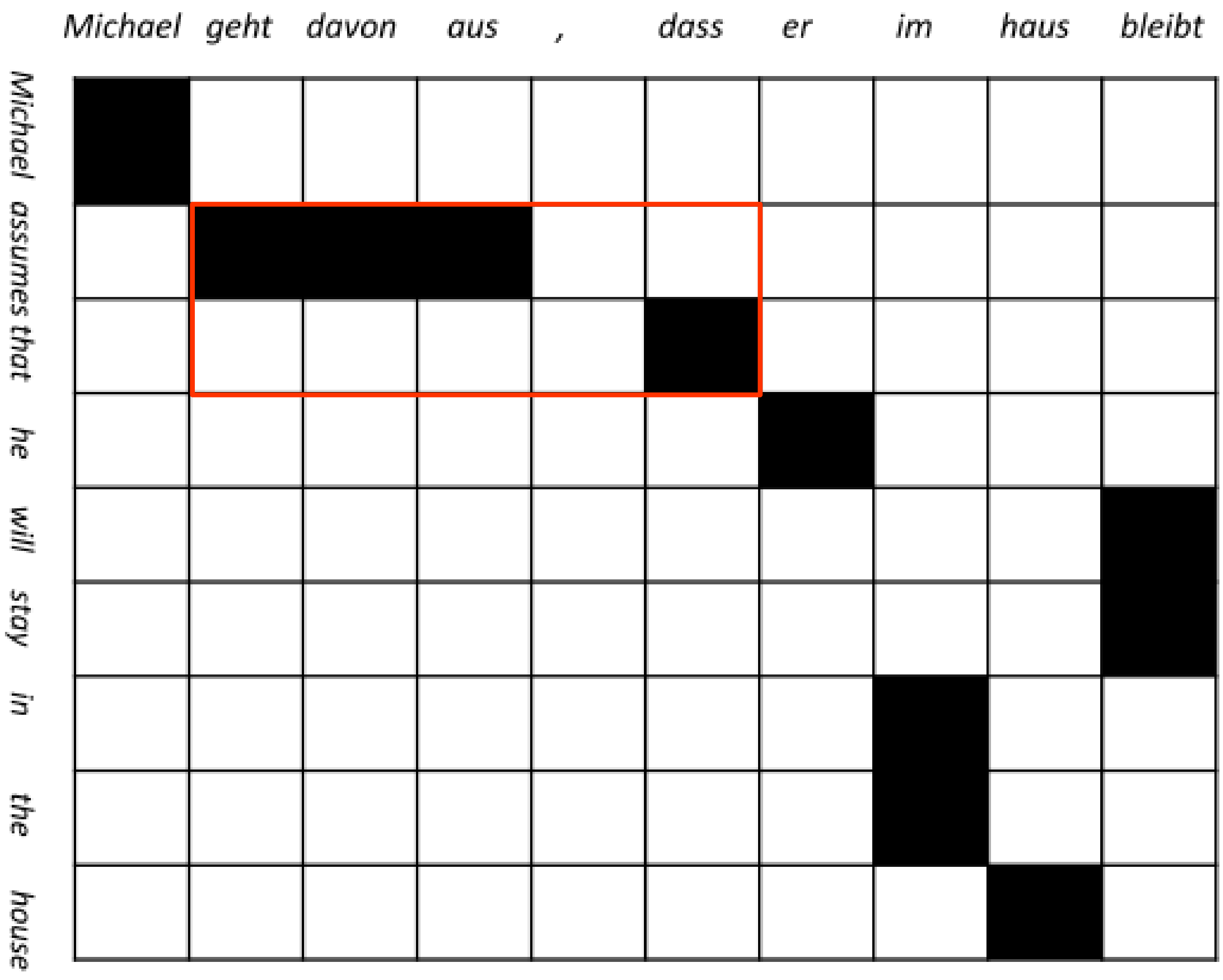
Phrase Pair Extraction



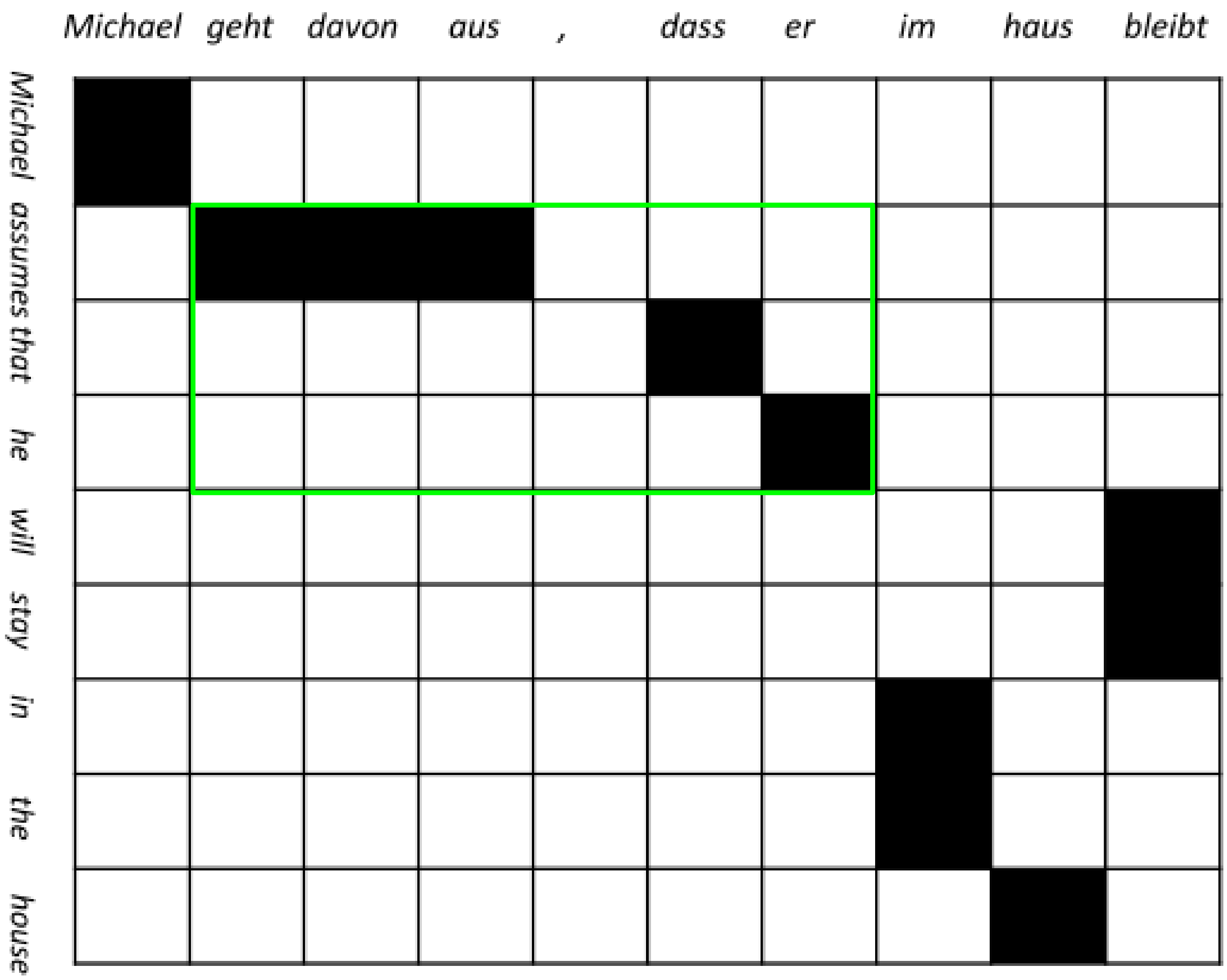
Phrase Pair Extraction



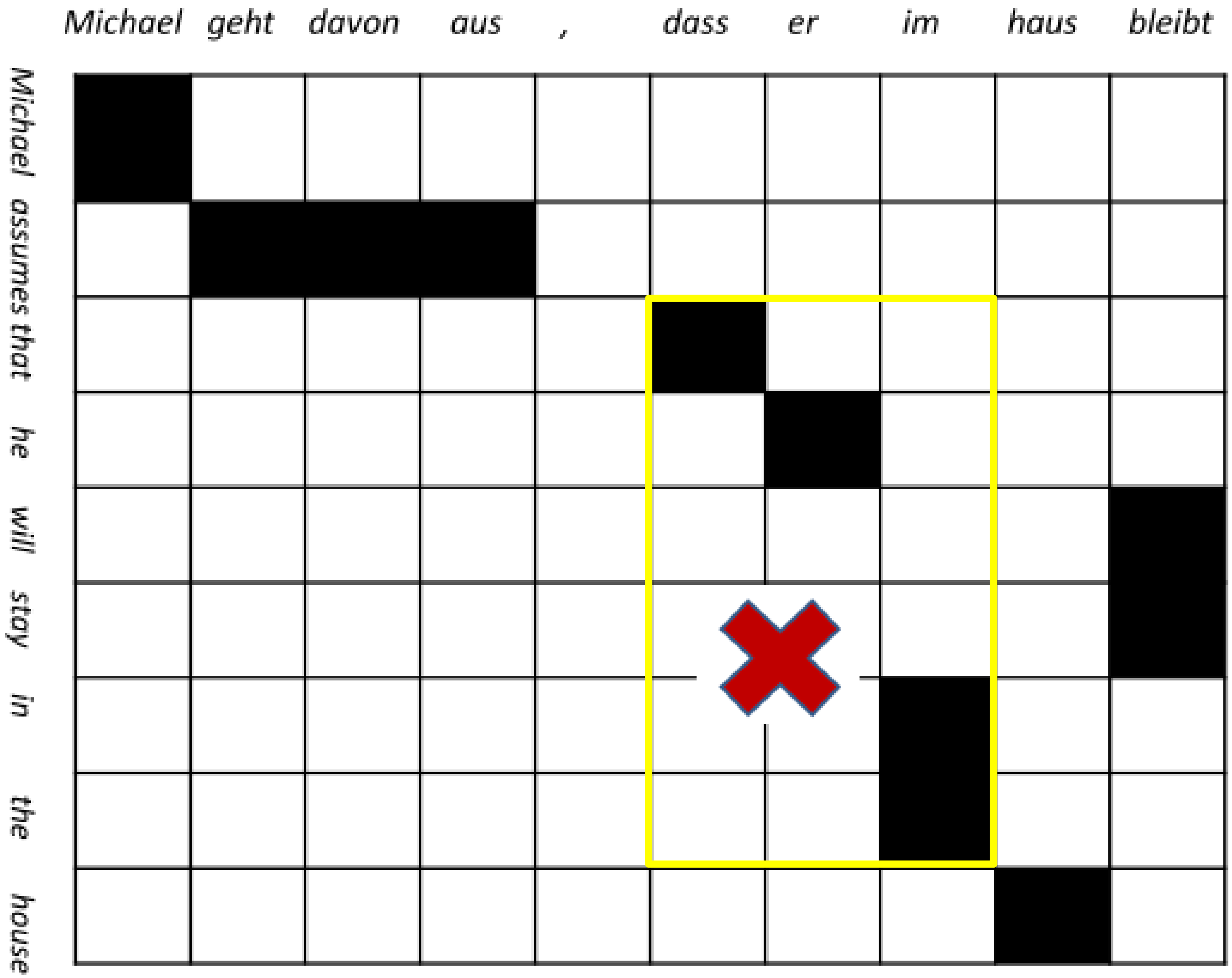
Phrase Pair Extraction



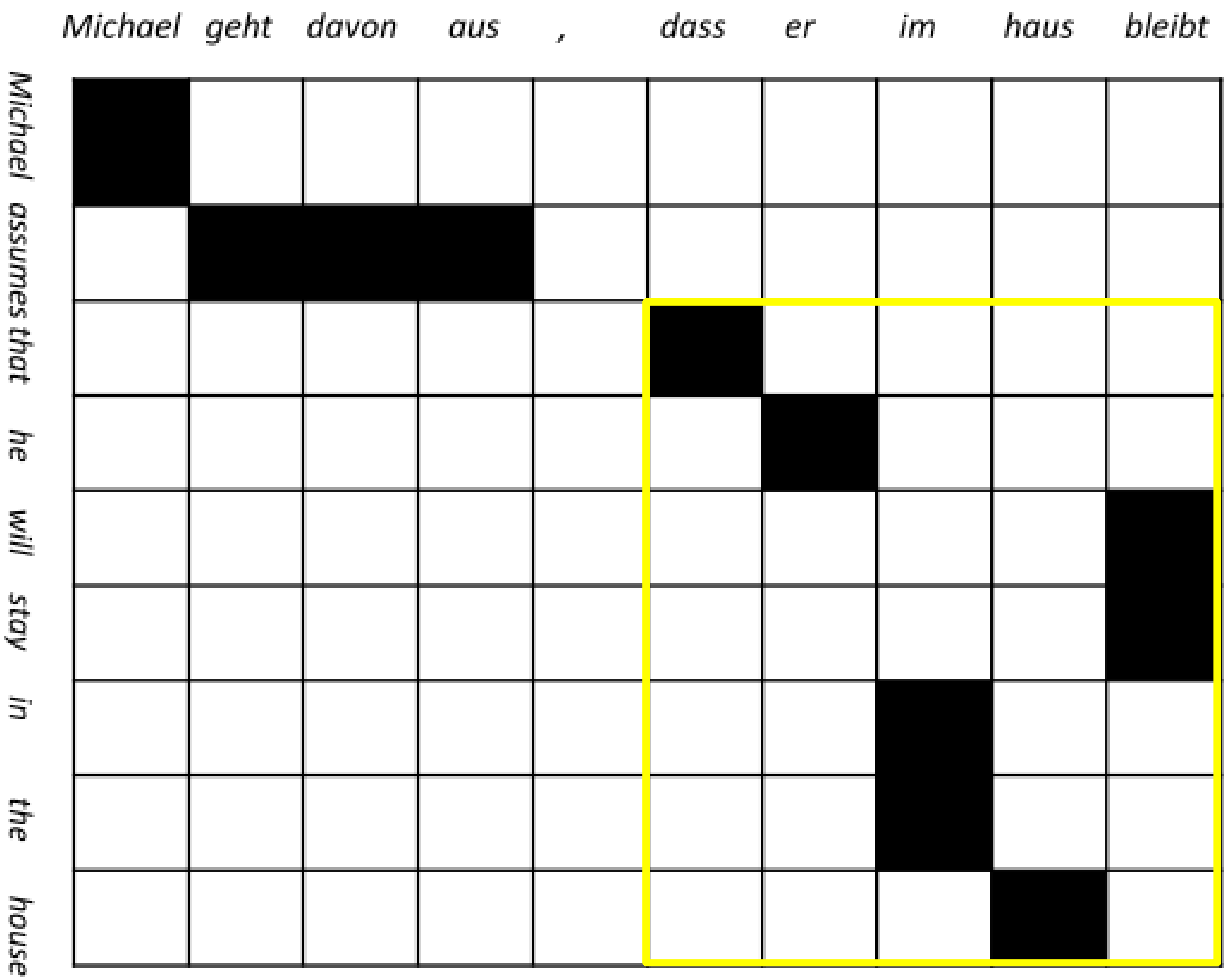
Phrase Pair Extraction



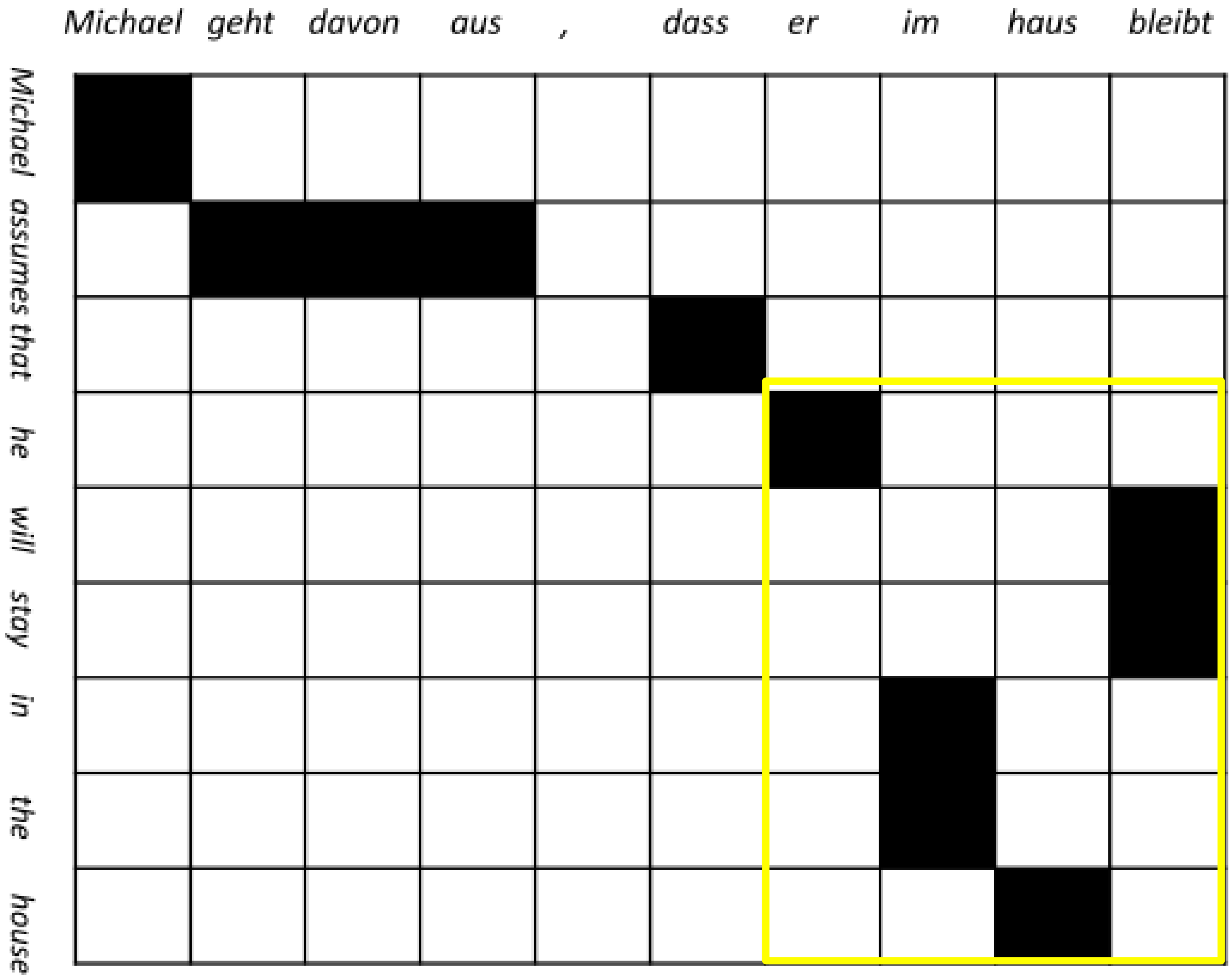
Phrase Pair Extraction



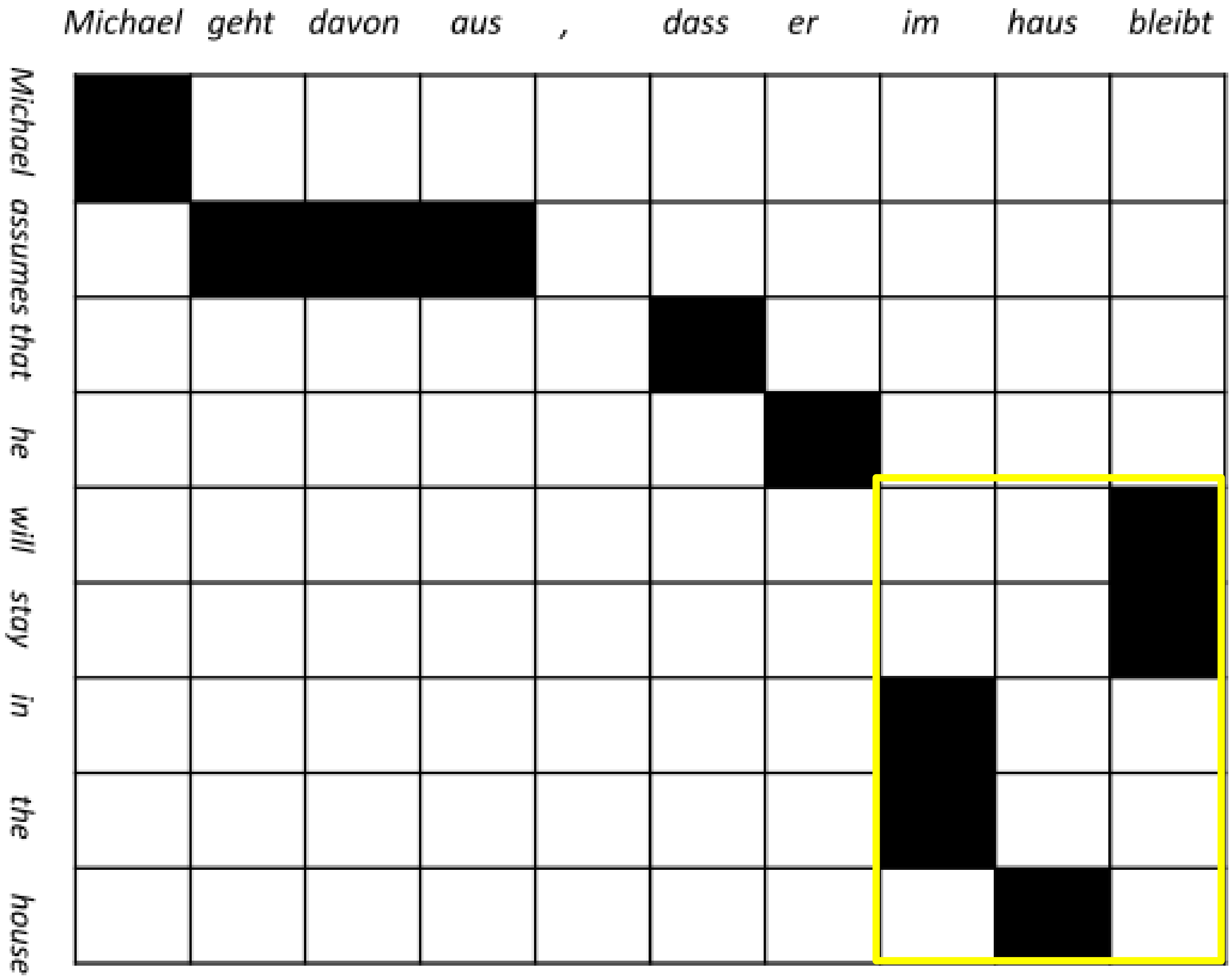
Phrase Pair Extraction



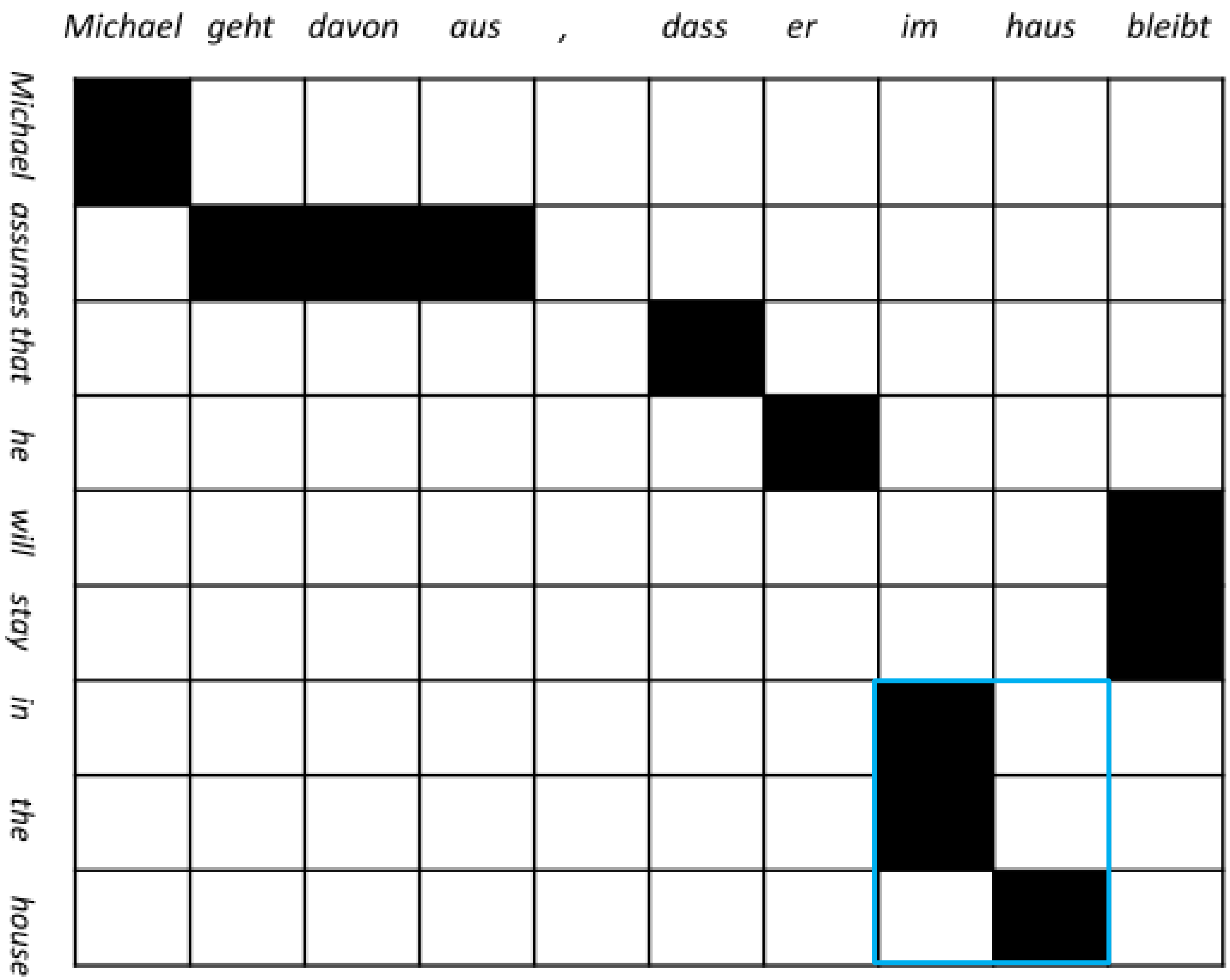
Phrase Pair Extraction



Phrase Pair Extraction



Phrase Pair Extraction



Phrase Pair Extraction (More)

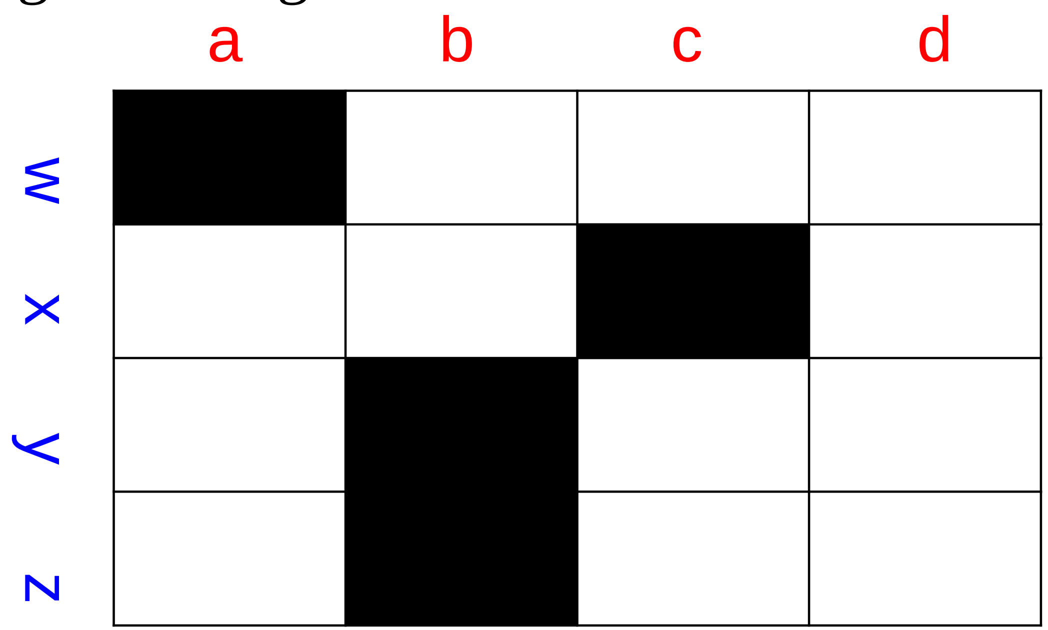
- michael assumes | michael geht davon aus / michael geht davon aus ,
- assumes that | geht davon aus , dass
- assumes that he | geht davon aus , dass er
- that he | dass er / , dass er
- in the house | im haus
- michael assumes that | michael geht davon aus , dass
- michael assumes that he | michael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

Exercise

Source: a b c d

Target: w x y z

Extract all bilingual phrase pairs consistent with the following word alignment.



Exercise

w | a

w x y z | a b c

w x y z | a b c d

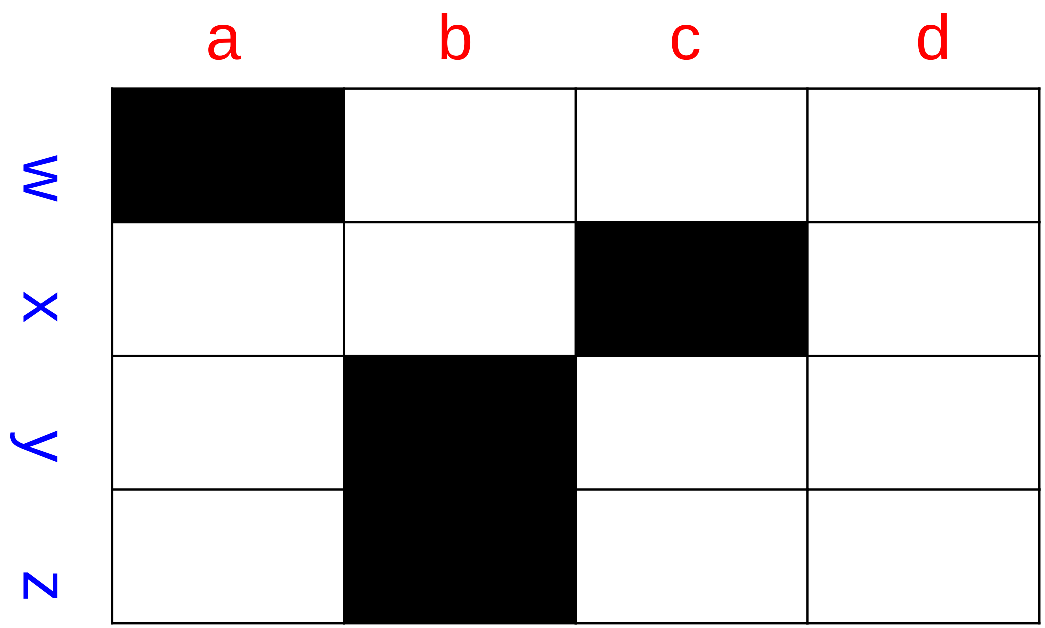
x | c

x | c d

x y z | b c

x y z | b c d

y z | b



Content

- Phrase-based Translation Model
- Learning a Phrase Translation Table
- Bidirectional Word Alignment
- Phrase Pair Extraction
- **Phrase Translation Probability**
- Exercises

Scoring Phrase Translations

- Phrase pair extraction: collect all phrase pairs from the data
- Phrase pair scoring: assign probabilities to phrase translations
- Score by relative frequency (MLE):

$$\phi(\bar{f}|\bar{e}) = \frac{count(\bar{e}, \bar{f})}{\sum_{\bar{f}_i} count(\bar{e}, \bar{f}_i)}$$

Scoring Phrase Translations

- Score by relative frequency (MLE) (the other direction):

$$\phi(\bar{e}|\bar{f}) = \frac{count(\bar{e}, \bar{f})}{\sum_{\bar{e}_i} count(\bar{f}, \bar{e}_i)}$$

Example

- Phrase translations for natuerlich, calculate the phrase translation probability.

Translation	Counts
of course	50
naturally	30
of course ,	15
, of course ,	10

Scoring Phrase Translations

$$\phi(\bar{e}|\bar{f}) = \frac{count(\bar{e},\bar{f})}{\sum_{\bar{e}_i} count(\bar{f},\bar{e}_i)}$$

$$\phi(\bar{e}|\bar{f}) = \phi(of\ course|natuerlich) = \frac{50}{50+30+15+10} = 0.5$$

$$\phi(\bar{e}|\bar{f}) = \phi(naturally|natuerlich) = \frac{30}{50+30+15+10} = 0.3$$

$$\phi(\bar{e}|\bar{f}) = \phi(of\ course\ ,|natuerlich) = \frac{15}{50+30+15+10} = 0.15$$

$$\phi(\bar{e}|\bar{f}) = \phi(,of\ course\ ,|natuerlich) = \frac{5}{50+30+15+10} = 0.05$$

Example: a Real Phrase Table

Source side: French

Target side: English

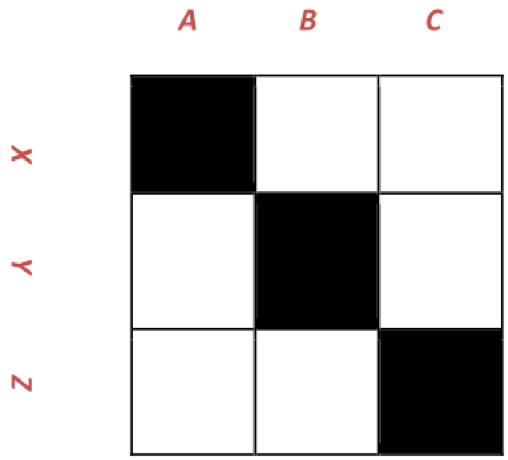
```
de l' immigration , ||| of immigration , ||| 0.5 0.0792945 1 0.0953929 |||
de l' immigration ||| immigration ||| 0.0769231 0.0872234 0.5 0.402069 |||
de l' immigration ||| of immigration ||| 0.5 0.10012 0.5 0.115455 ||| 0-0
de l' immobilier amé ricain ||| us housing ||| 0.5 0.00182555 1 0.0649596
de l' immobilier pour ||| of housing in ||| 1 0.000297943 1 0.00156694 |||
de l' immobilier ||| housing ||| 0.0769231 0.0173907 0.25 0.16069 ||| 1-0
de l' immobilier ||| of housing ||| 0.5 0.0199621 0.25 0.0461423 ||| 0-0 1
de l' immobilier ||| real estate ||| 0.333333 0.0447481 0.25 0.0201379 |||
de l' immobilier ||| remain ||| 0.05 0.000692525 0.25 0.04 ||| 2-0 ||| 20
```

Content

- Phrase-based Translation Model
- Learning a Phrase Translation Table
- Bidirectional Word Alignment
- Phrase Pair Extraction
- Phrase Translation Probability
- **Exercises**

Exercise 1

- List all phrase pairs that are consistent with the following word alignment:



Solution 1

X | A

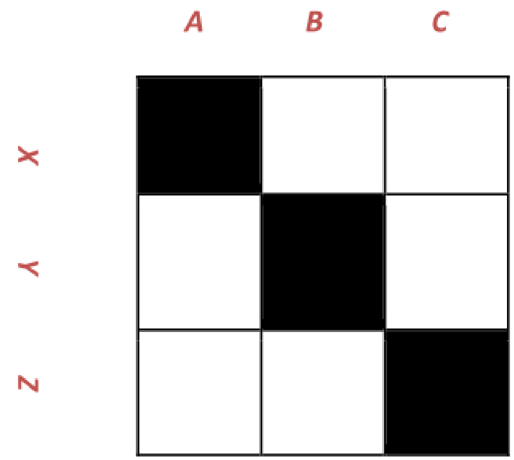
X Y | A B

X Y Z | A B C

Y | B

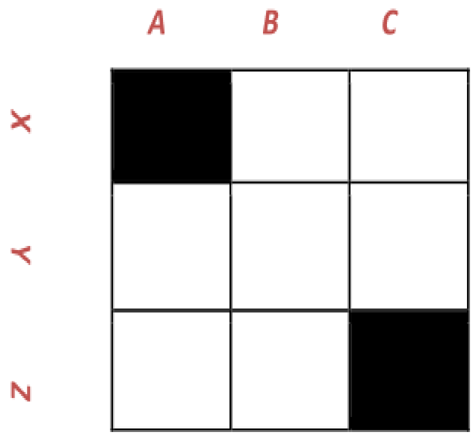
Y Z | B C

Z | C



Exercise 2

- List all phrase pairs that are consistent with the following word alignment:



Solution 2

X | A

X | A B

X Y | A

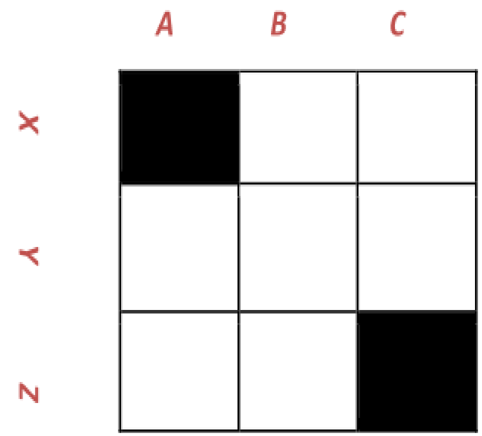
X Y | A B

X Y Z | A B C

Z | C

Y Z | C

Y Z | B C



Exercise 3

- Given the following statistics of extracted bilingual phrases in terms of Chinese phrase "xihuan paobu", please calculate the translation probabilities for each phrase pair.

Translation	Counts
likes running	1500
like running	800
likes jogging	700
love running	100

Solution 3

$$\phi(likes\ running|xihuan\ paobu) = \frac{1500}{1500 + 800 + 700 + 100} = 0.484$$

$$\phi(like\ running|xihuan\ paobu) = \frac{800}{1500 + 800 + 700 + 100} = 0.258$$

$$\phi(likes\ running|xihuan\ paobu) = \frac{700}{1500 + 800 + 700 + 100} = 0.226$$

$$\phi(likes\ running|xihuan\ paobu) = \frac{100}{1500 + 800 + 700 + 100} = 0.032$$

Translation	Counts
likes running	1500
like running	800
likes jogging	700
love running	100

Discussion