数据科学与计算机学院
School of Data and Computer Science

自然语言处理

Natural Language Processing

权小军 教授

中山大学数据科学与计算机学院

quanxj3@mail.sysu.edu.cn

八、机器翻译

基本翻译方法

- 直接转换法
- 基于规则的翻译方法
- 基于中间语言的翻译方法
- 基于语料库的翻译方法
  - 基于事例的翻译方法
  - 统计翻译方法
  - 神经网络机器翻译

基本翻译方法(一)：直接转换法

❑ 直接转换法

从源语言句子的表层出发，将单词、短语或句子直接置换成目标语言译文，必要时进行简单的词序调整。这类翻译系统一般针对某一个特定的语言对，将分析与生成、语言数据、文法和规则与程序等都融合在一起。例如：

I like Mary. → Me(I) gusta(like) Maria(Mary).

基本翻译方法(二)：基于规则

❑ 基于规则的翻译方法(Rule-based)

对源语言和目标语言均进行适当描述、把翻译机制与语法分开、用规则描述语法的实现思想，这就是基于规则的翻译方法。

基本翻译方法(二)：基于规则

基于规则的翻译过程分成6个步骤：

(a) 对源语言句子进行词法分析

(b) 对源语言句子进行句法/语义分析

(c) 源语言句子结构到译文结构的转换

(d) 译文句法结构生成

(e) 源语言词汇到译文词汇的转换

(f) 译文词法选择与生成

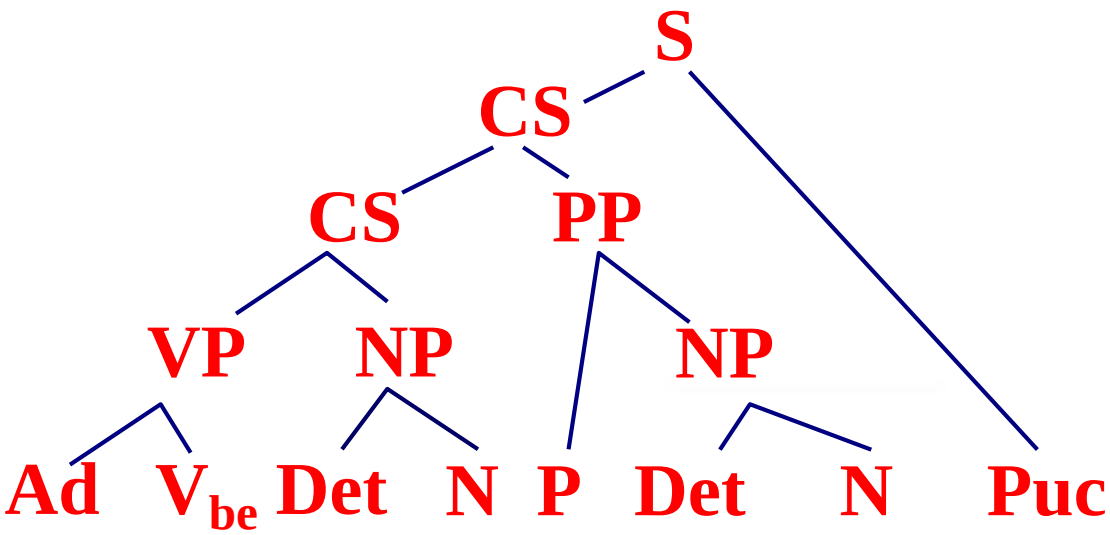
基本翻译方法(二)：基于规则

给定源语言句子： There is a book on the desk.

■Step 1: 词法分析：

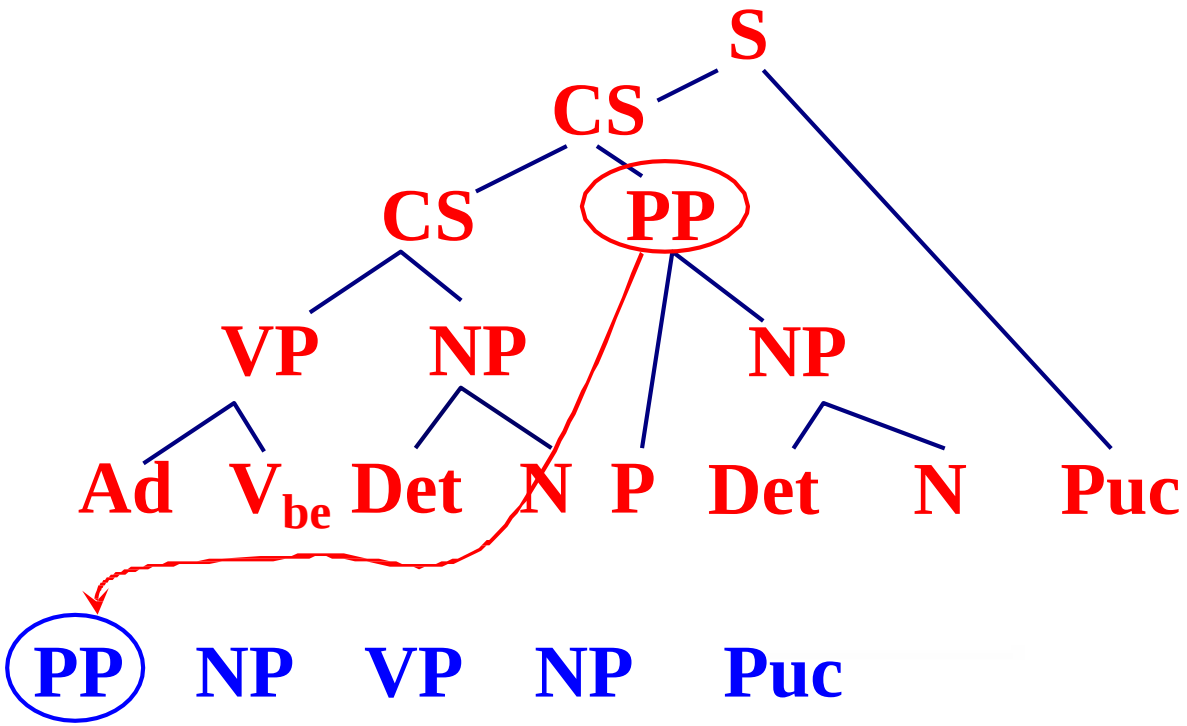
There/Ad is/V_{be} a/Det book/N on/P the/Det desk/N ./Puc

■Step 2: 利用句法规则进行句法结构分析：



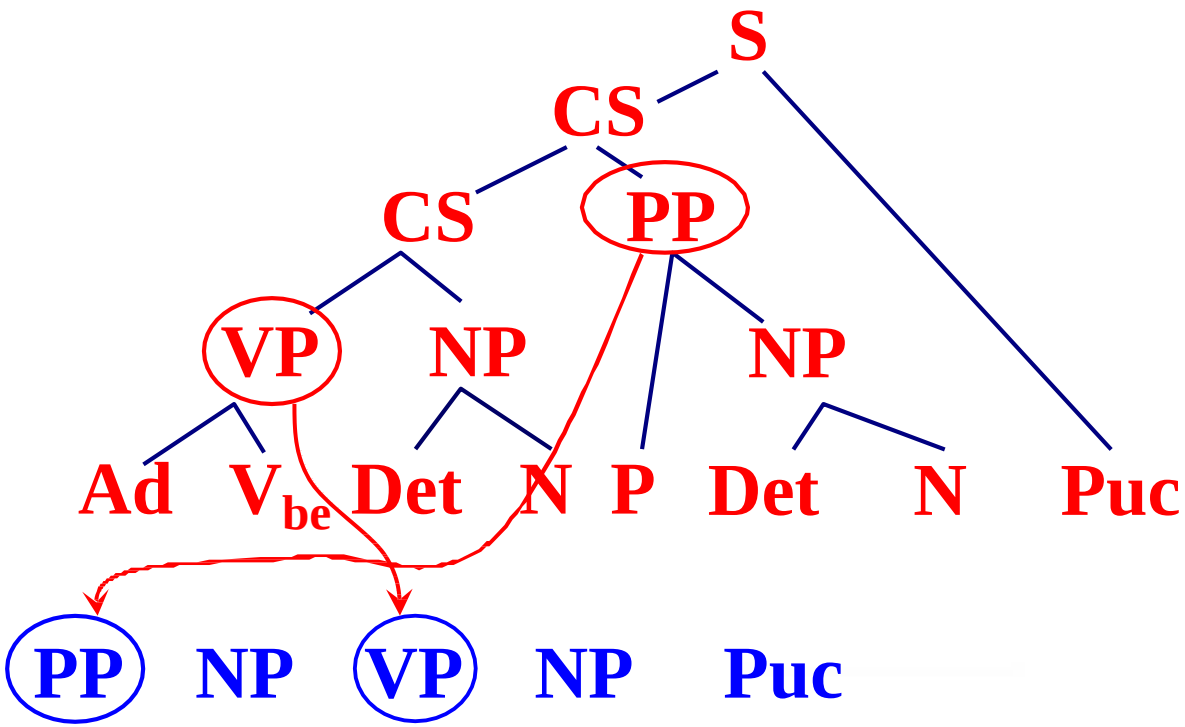
基本翻译方法(二)：基于规则

■Step 3: 利用转换规则将源语言句子结构转换成目标语言句子结构



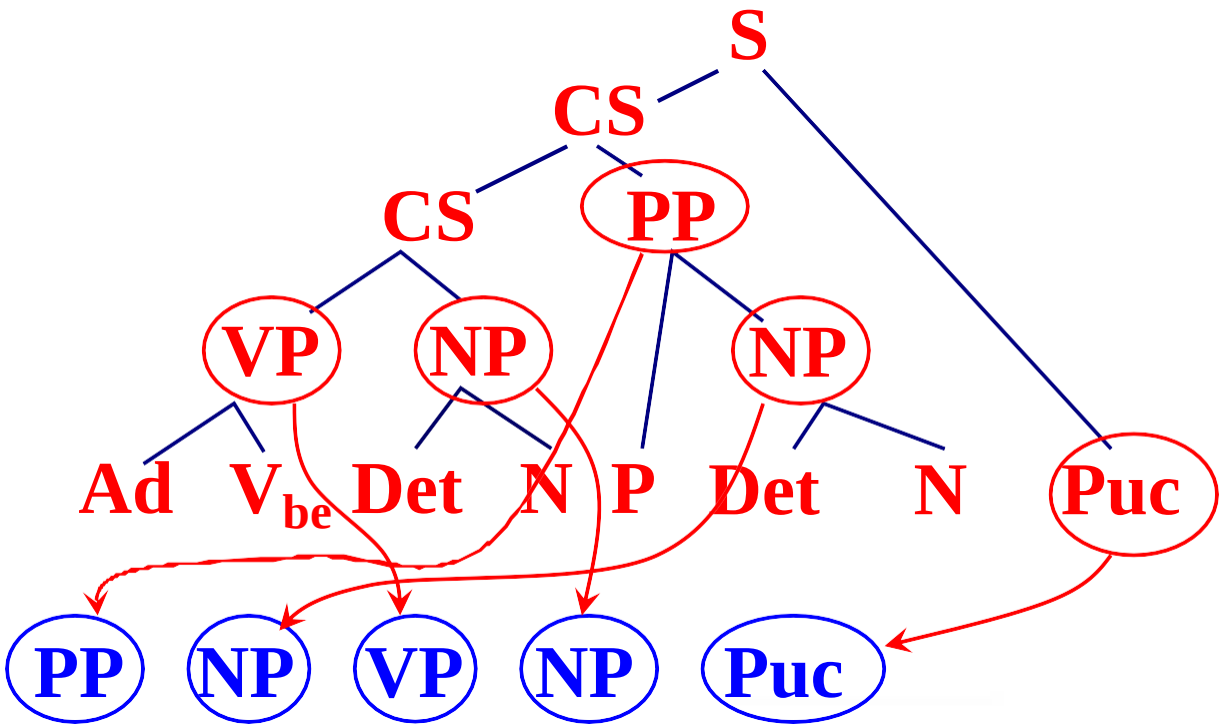
基本翻译方法(二)：基于规则

■Step 3: 利用转换规则将源语言句子结构转换成目标语言句子结构



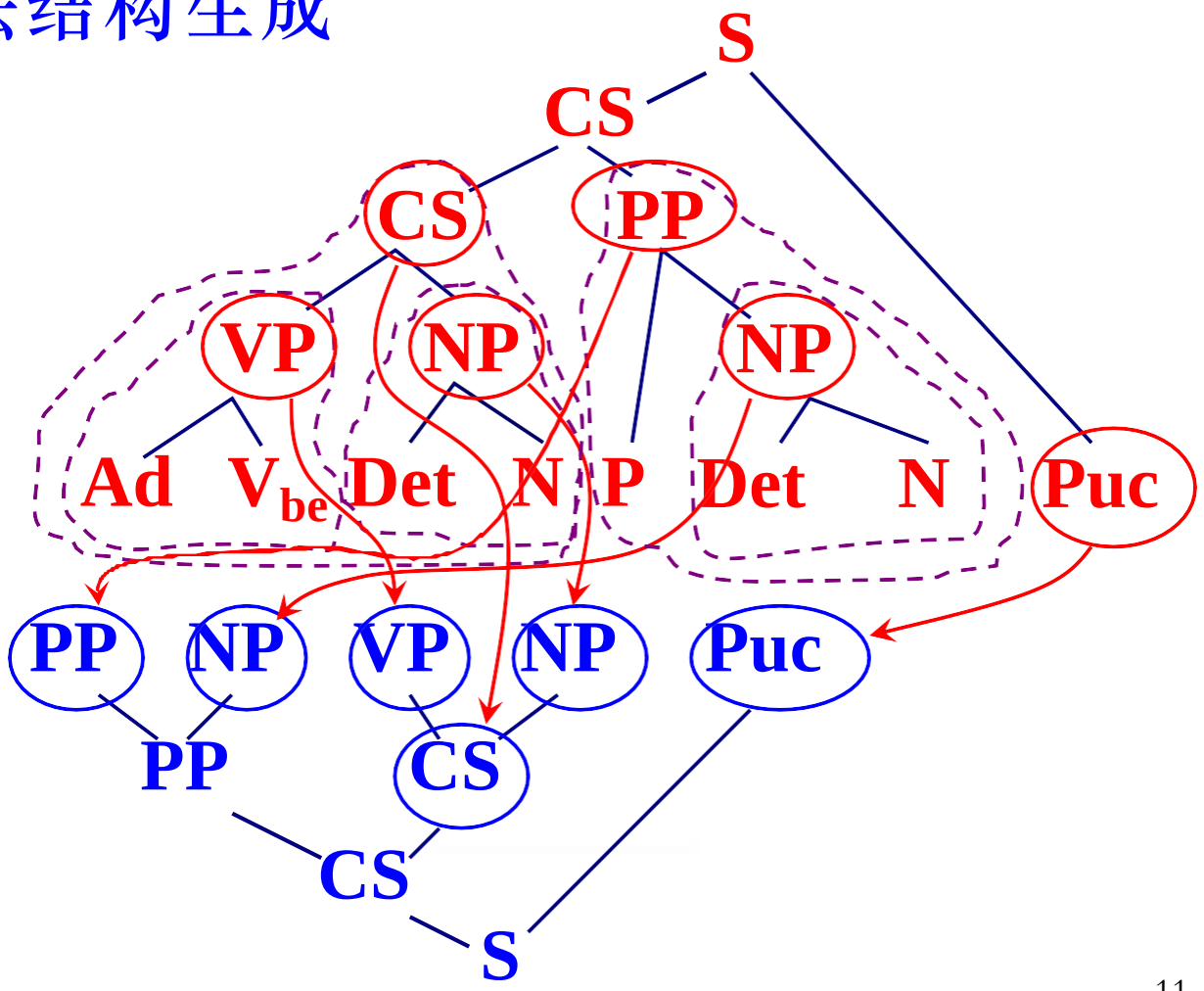
基本翻译方法(二)：基于规则

■Step 3: 利用转换规则将源语言句子结构转换成目标语言句子结构



基本翻译方法(二)：基于规则

■Step 4:译文句法结构生成



基本翻译方法(二)：基于规则

- Step 5: 将源语言词汇翻译成目标语言词汇

# there	Ad: 在那里
# be	V_{be}: 是
# there be	VP: 在…有
# a	Det: 一，一个，一本…
# book	N: 书，书籍; V: 预订

- Step 6: 译文词法处理和目标语言句子生成：

  在桌子上有一本书。

基本翻译方法(二)：基于规则

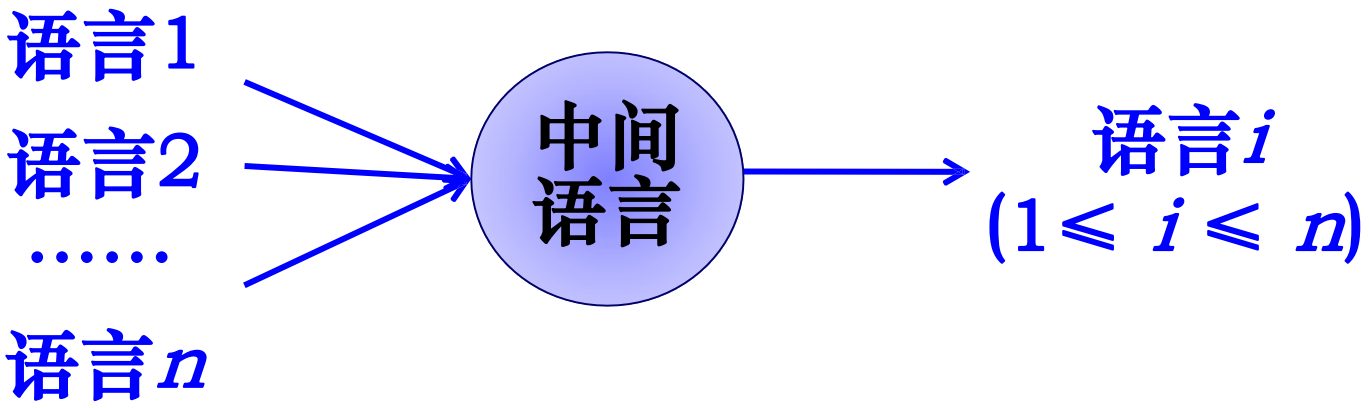
❑ 对基于规则的翻译方法的评价：

- 优点：可以较好地保持原文的结构，产生的译文结构与源文的结构关系密切，尤其对于语言现象已知的或句法结构规范的源语言语句具有较强的处理能力和较好的翻译效.
- 弱点：规则一般由人工编写，工作量大，主观性强，一致性难以保障，不利于系统扩充，对非规范语言现象缺乏相应的处理能力.

基本翻译方法(三)：基于中间语言

❑ 基于中间语言的翻译方法

- 方法：输入语句→中间语言→ 翻译结果



基本翻译方法(三)：基于中间语言

❑ 关于中间语言的定义

- 国际先进语音翻译研究联盟(C-STAR)定义的中间转换格式(Interchange Format)
- 日本东京联合国大学(United Nations University)提出的通用网络语言(Universal Networking Language)

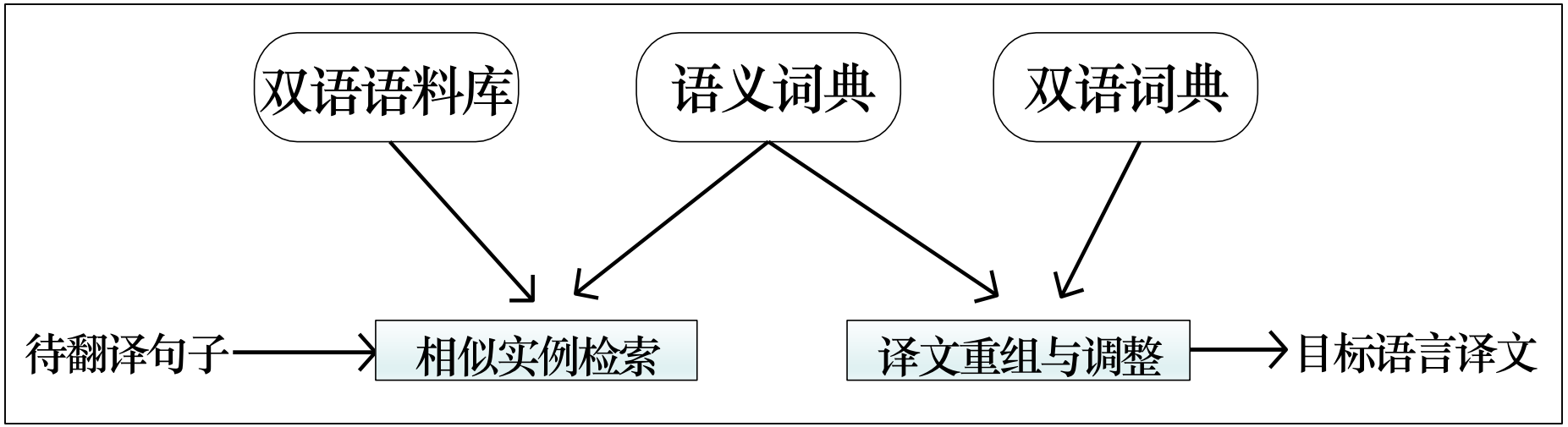
基本翻译方法(三)：基于中间语言

❑ 对基于中间语言的翻译方法评价：

- 优点：中间语言的设计可以不考虑具体的翻译语言对，因此，该方法尤其适合多语言之间的互译。
- 弱点：如何定义和设计中间语言的表达方式，以及如何维护并不是一件容易的事情，中间语言在语义表达的准确性、完整性等很多方面，都面临若干困难。

基本翻译方法(四)：基于事例

- 基于事例(实例)的翻译方法(Example-based)
  - 方法：输入语句→与事例相似度比较→ 翻译结果
  - 资源：大规模事例库



基本翻译方法(四)：基于事例

❑ 对基于实例的翻译方法评价：

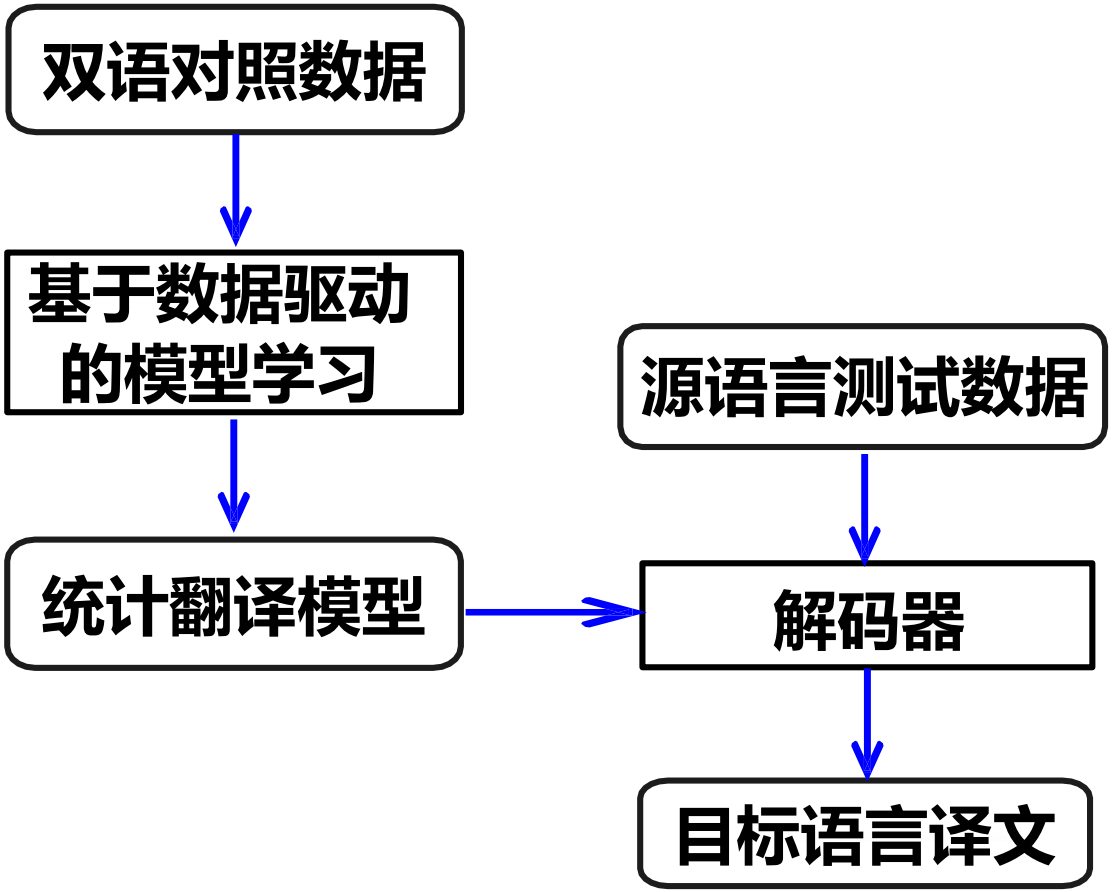
- 优点：不要求源语言句子必须符合语法规定，翻译机制一般不需要对源语言句子做深入分析.
- 弱点：两个不同的句子之间的相似性往往难以把握； 系统往往难以处理事例库中没有记录的陌生的语言现象，而且当事例库达到一定规模时，其事例检索的效率较低.

基本翻译方法

❑ 其它翻译方法

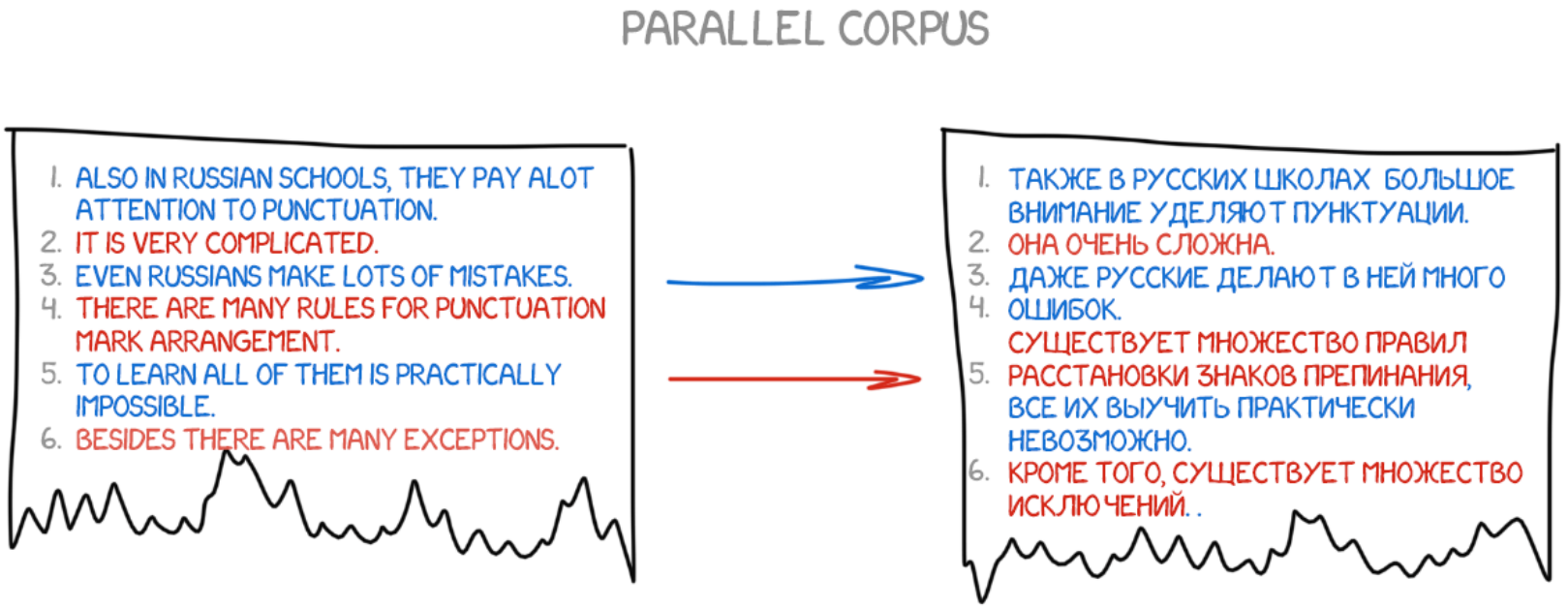
- ❖ 统计翻译方法(statistical method)
- ❖ 基于神经网络(neural network)的翻译方法

统计翻译的思想



统计翻译的思想

❑ 平行语料



统计翻译基本原理

❑ 噪声信道模型

一种语言T 由于经过一个噪声信道而发生变形，从而在信道的另一端呈现为另一种语言S（信道意义上的输出，翻译意义上的源语言)。翻译问题实际上就是如何根据观察到的S，恢复最为可能的T 问题.

统计翻译基本原理

- 源语言句子： $S = s_1^m = s_1 s_2 \cdots s_m$
- 目标语言句子： $T = t_1^l = t_1 t_2 \cdots t_n$
- 贝叶斯公式： $P(T|S) = \dfrac{P(T) \times P(S|T)}{P(S)}$

$$T' = \underset{T}{\operatorname{argmax}} P(T) \times P(S|T)$$

语言模型
Language model, LM

翻译模型
Translation model, TM

统计翻译基本原理

统计翻译中的三个关键问题：

(1) 估计语言模型概率 $p(T)$；

(2) 估计翻译概率 $p(S|T)$；

(3)快速有效地搜索T使得 $p(T) \times p(S|T)$最大;

统计翻译基本原理

- 估计语言模型概率$p(T)$

给定句子：$T = t_1^l = t_1 t_2 \cdots t_n$

怎么估算T的概率？？？

句子概率：$P(T) = P(t_1)P(t_2|t_1) \cdots P(t_n|t_1 t_2 \cdots t_{n-1})$

基于n-gram 来计算！

统计翻译基本原理

- 翻译概率 $p(S|T)$ 的计算

$$P(S|T)=\sum_{A} P(S,A|T)$$

$$\Rightarrow \quad P(S,A|T) \quad ???$$

$$P(S,A|T)=p(m|T)\times P(A|T,m)\times P(S|T,A,m)$$

对位模型

词汇翻译模型

统计翻译基本原理

$$P(S,A|T) = p(m|T) \times P(A|T,m) \times P(S|T,A,m)$$

$$= p(m|T) \prod_{j=1}^{m} p\left(a_j|a_1^{j-1}, s_1^{j-1}, m, T\right) \times p\left(s_j|a_1^{j}, s_1^{j-1}, m, T\right)$$

基于上式，IBM 的研究人员通过采用不同的假设条件得到了5个翻译模型，分别称作 IBM 翻译模型1、2、3、4 和 5

IBM 翻译模型1

$$p(m|T)\prod_{j=1}^{m}p\left(a_j|a_1^{j-1},s_1^{j-1},m,T\right)\times p\left(s_j|a_1^{j},s_1^{j-1},m,T\right)$$

翻译模型1：

(1) 假设 $\varepsilon\equiv p(m|T)$是一个较小的常量；

(2) 假设 $a_j\sim uniform\ (0,1,2,\cdots,n)$ $p\left(a_j|a_1^{j-1},s_1^{j-1},m,T\right)=\frac{1}{n+1}$

(3) 假设 $s_j\sim Categorical\left(\theta_{t_{a_j}}\right)$ $p\left(s_j|a_1^{j},s_1^{j-1},m,T\right)=p\left(s_j|t_{a_j}\right)$

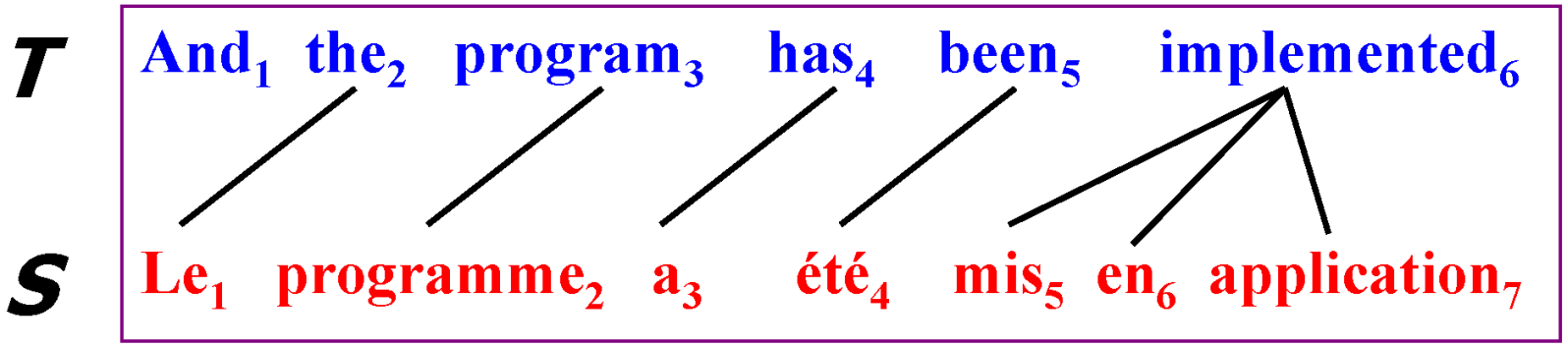
IBM 翻译模型1

$$P(S,A|T) = p(m|T)\prod_{j=1}^{m} p\left(a_j|a_1^{j-1}, s_1^{j-1}, m, T\right) \times p\left(s_j|a_1^{j}, s_1^{j-1}, m, T\right)$$

$$= \varepsilon \prod_{j=1}^{m} \frac{1}{n+1} \times p\left(s_j|t_{a_j}\right)$$

$$= \frac{\varepsilon}{(n+1)^m} \prod_{j=1}^{m} p\left(s_j|t_{a_j}\right)$$

IBM 翻译模型1



$$P(S,A|T) == \frac{\varepsilon}{(l+1)^m}\prod_{j=1}^{m} p\left(s_j|t_{a_j}\right)$$

$$\frac{\varepsilon}{(6+1)^7} \times [p(Le|the) \times \cdots \times p(application|implemented)]$$

IBM 翻译模型1

$$P(S,A|T)=\frac{\varepsilon}{(l+1)^m}\prod_{j=1}^{m}p\left(s_j|t_{a_j}\right)$$

$$P(S|T)=\sum_{A}P(S,A|T)=\sum_{A}\frac{\varepsilon}{(l+1)^m}\prod_{j=1}^{m}p\left(s_j|t_{a_j}\right)$$

$$=\frac{\varepsilon}{(l+1)^m}\sum_{a_{1=0}}^{n}\cdots\sum_{a_m=0}^{n}\prod_{j=1}^{m}p\left(s_j|t_{a_j}\right)$$

如何训练？

IBM 翻译模型1

$$argmax\ P(S|T)$$

$$w.r.t.\ \sum_{s} p(s|t) = 1$$

$$h(p,\lambda) = P(S|T) - \sum_{t} \lambda_t \left(\sum_{s} p(s|t) - 1 \right)$$

$$= \frac{\varepsilon}{(l+1)^m} \sum_{a_1=0}^{n} \cdots \sum_{a_m=0}^{n} \prod_{j=1}^{m} p\left(s_j|t_{a_j}\right) - \sum_{t} \lambda_t \left(\sum_{s} p(s|t) - 1 \right)$$

IBM 翻译模型1

$$h(p,\lambda) =$$

$$= \frac{\varepsilon}{(l+1)^m} \sum_{a_1=0}^{n} \cdots \sum_{a_m=0}^{n} \prod_{j=1}^{m} p\left(s_j|t_{a_j}\right) - \sum_t \lambda_t \left(\sum_s p(s|t) - 1 \right)$$

$$\frac{\partial h(p,\lambda)}{\partial p(s|t)} = 0$$

$$\frac{\varepsilon}{(l+1)^m} \sum_{a_1=0}^{n} \cdots \sum_{a_{m=0}}^{n} \sum_{j=1}^{m} \delta(s=s_j)\delta\left(t=t_{a_j}\right) \frac{1}{p(s|t)} \prod_{k=1}^{m} p(s_k|t_{a_k})$$

$$- \lambda_t = 0$$

IBM 翻译模型1

$$\frac{\varepsilon}{(l+1)^m}\sum_{a_1=0}^{n}\cdots\sum_{a_m=0}^{n}\sum_{j=1}^{m}\delta(s=s_j)\delta\left(t=t_{a_j}\right)\frac{1}{p(s|t)}\prod_{k=1}^{m}p\left(s_k|t_{a_k}\right)$$
$$-\lambda_t=0$$

$$p(s|t)=$$
$$\frac{1}{\lambda_t}\times\frac{\varepsilon}{(l+1)^m}\sum_{a_1=0}^{n}\cdots\sum_{a_m=0}^{n}\sum_{j=1}^{m}\delta(s=s_j)\delta\left(t=t_{a_j}\right)\prod_{k=1}^{m}p\left(s_k|t_{a_k}\right)$$

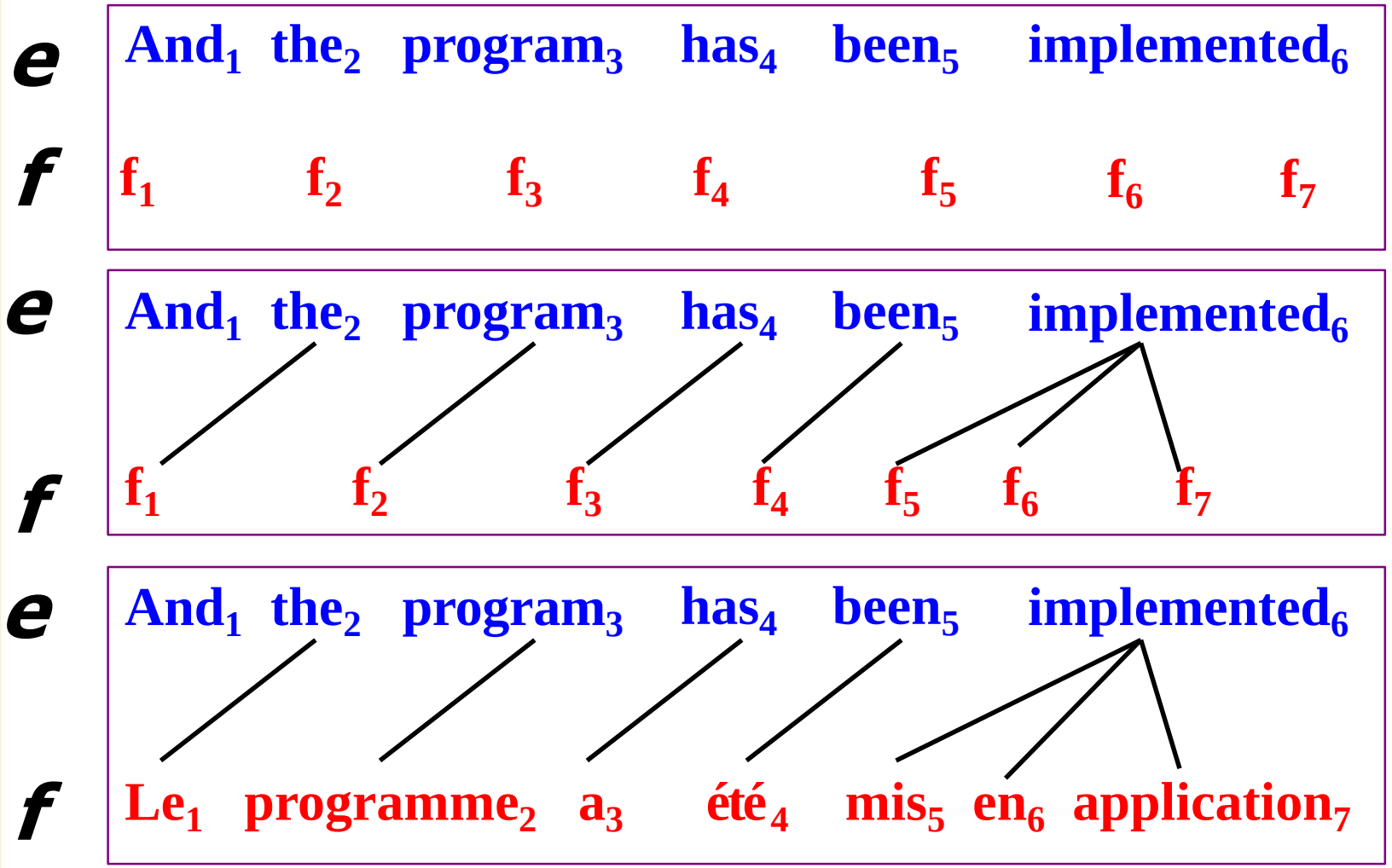
IBM 翻译模型1

忽略详细的数学推导，IBM 翻译模型1表示为如下等式：

$$P(S|T) = \frac{\varepsilon}{(n+1)^m} \sum_{a_1=0}^{n} \cdots \sum_{\iota_m=0}^{n} \prod_{j=1}^{m} p\left(s_j|t_{a_j}\right)$$

$$= \frac{\varepsilon}{(n+1)^m} \prod_{j=1}^{m} \sum_{a_j=0}^{n} p\left(s_j|t_{a_j}\right)$$

IBM 翻译模型1



$$\varepsilon \equiv p(m|T)$$

$$\frac{1}{n+1}$$

$$p\left(s_j|t_{a_j}\right)$$

IBM 翻译模型2

在IBM 模型2中，除了假定概率 $P(a_j \mid a_1^{j-1}, s_1^{j-1}, m, T)$ 依赖于位置 j、对位关系a_j 和源语言句子长度m以及目标语言句子长度n以外，另外两个假设与IBM模型1一样。

引入了对位概率(alignment probabilities)**的概念：**

$$a(a_j|j,m,n) = P(a_j|a_1^{j-1}, s_1^{j-1}, m, n)$$

IBM 翻译模型2

对于每一个三元组(j, m, n)，对位概率满足如下约束条件：

$$\sum_{i=0}^{n} a(i \mid j, m, n) = 1$$

类似于IBM模型1的推导，得到模型2：

$$p(S \mid T) = \varepsilon \prod_{j=1}^{m} \sum_{i=0}^{n} p(s_j \mid t_i) \times a(i \mid j, m, n)$$

IBM 翻译模型2

对于每一个三元组(j, m, n)，对位概率满足如下约束条件：

$$\sum_{i=0}^{n} a(i \mid j, m, n) = 1$$

类似于IBM模型1的推导，得到模型2：

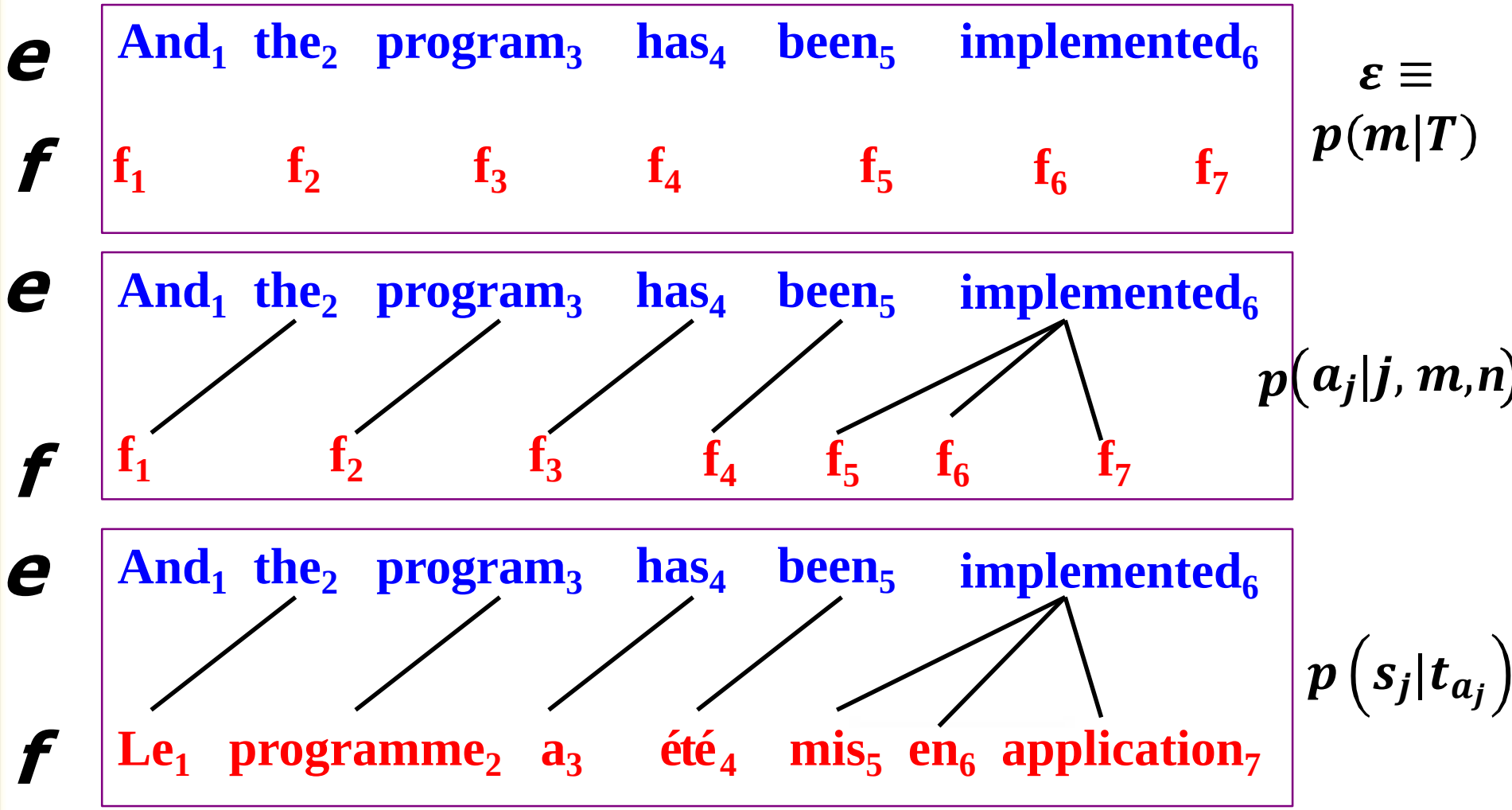
$$p(S \mid T) = \varepsilon \prod_{j=1}^{m} \sum_{i=0}^{n} p(s_j \mid t_i) \times a(i \mid j, m, n) \qquad (7)$$

$a(i \mid j, m, n)$ 对位概率设为常数，即 $\frac{1}{n+1}$

如果对位概率设为常数，IBM 模型2退化为模型1，即模型1是模型2的特例。

$$p(S \mid T) = \frac{\varepsilon}{(l+1)^m} \prod_{j=1}^{m} \sum_{i=0}^{l} p(s_j \mid t_i)$$

IBM 翻译模型2


e
And1 the2 program3 has4 been5 implemented6
f
f1 f2 f3 f4 f5 f6 f7
ε ≡ p(m|T)
e
And1 the2 program3 has4 been5 implemented6
f
f1 f2 f3 f4 f5 f6 f7
p(aj|j, m,n)
e
And1 the2 program3 has4 been5 implemented6
f
Le1 programme2 a3 été4 mis5 en6 application7
p(sj|taj)

基于短语的翻译模型

❑ 翻译基本单元由词转向短语

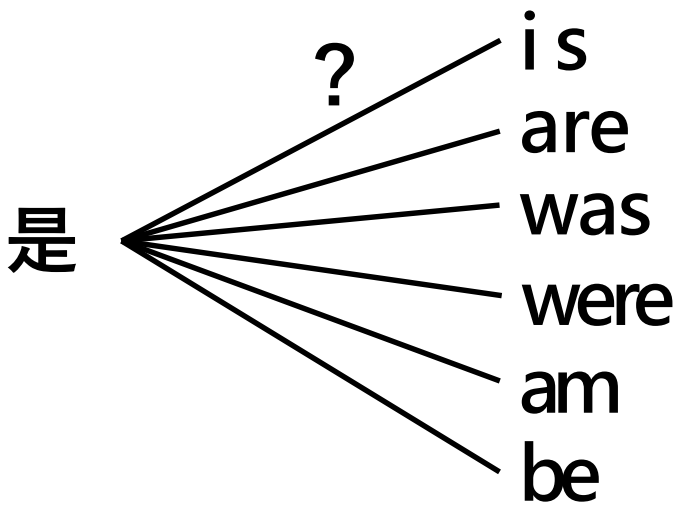
- 基于词的翻译模型的问题：
  - 很难处理词义消岐问题
  - 很难处理一对多、多对一和多对多的翻译问题

澳洲 是 与 北韩 有 邦交 的 少数 国家 之一

基于短语的翻译模型

- 词义消岐问题

澳洲 是 与 北韩 有 邦交 的 少数 国家 之一



在基于词的模型中，处理词义消岐问题需要充分利用上下文信息，并对上下文信息进行有效建模。

基于短语的翻译模型

- 一对多、多对一与多对多翻译问题

澳洲 是 与 北韩 有 邦交 的 少数 国家 之一

北 韩 → North Korea

邦 交 → the diplomatic relations

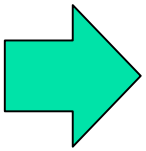
在基于词的模型中，处理上述问题同时需要准确的词汇翻译模型以及对位模型。

基于短语的翻译模型

- 以短语为基本翻译单元

澳洲 是 与 北韩 有 邦交 的 少数 国家 之一

短语翻译规则



(澳洲 是，Australia is)

(北韩，Korea North)

(邦交，the diplomatic relations)

基于短语的翻译模型

- 以短语为基本翻译单元

澳洲 是 与 北韩 有 邦交 的 少数 国家 之一

(澳洲 是， Australia is)

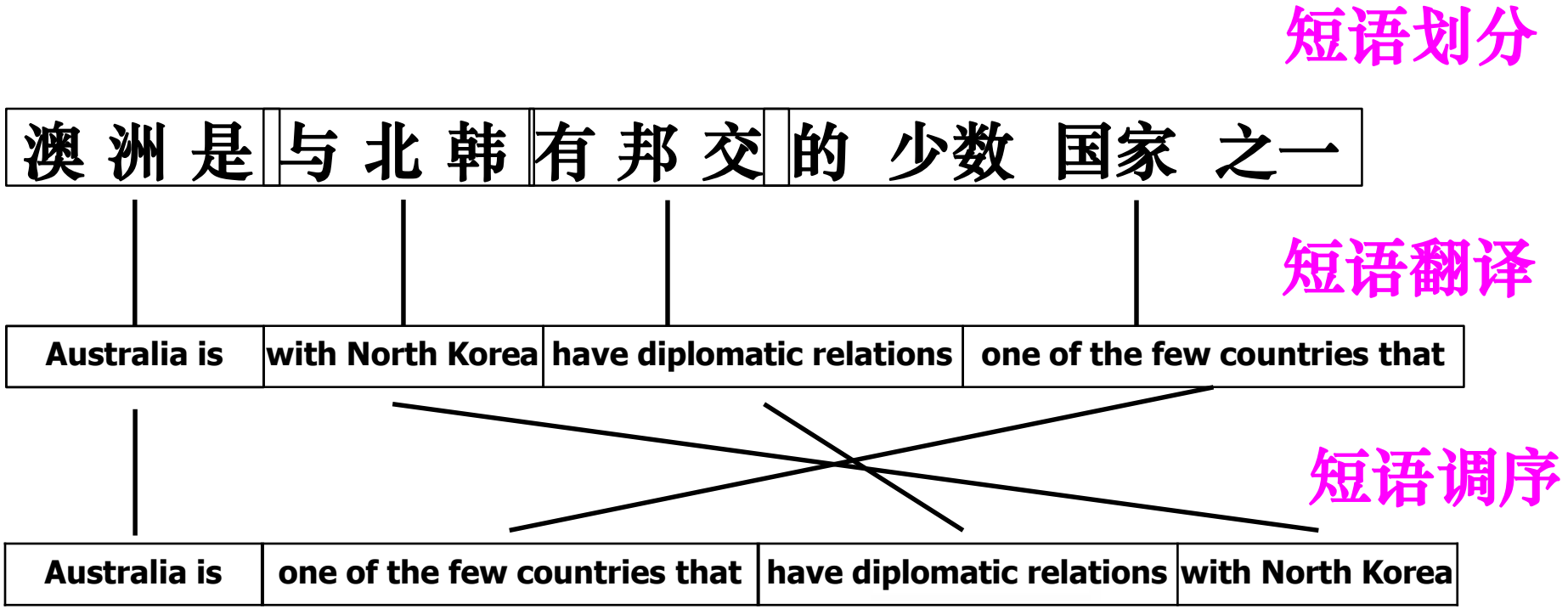
(与 北韩， with North Korea)

(有 邦交， have the diplomatic relations)

(的 少数 国家 之一， one of the few countries that)

基于短语的翻译模型

- 以短语为基本翻译单元



基于短语的翻译模型

- **基于短语的翻译模型** [Koehn, 2003]

  短语：连续的词串（非句法意义）

- 注意：这里所说的短语指一个连续的词串(*n*-gram)，不一定是语言学中定义的短语(phrase)，如：

  我想预订一个单人间。

  I would like to reserve a single room.

基于短语的翻译模型

$$T' = \arg\max_{T} P(T|S)$$

$$= \arg\max_{T, S_1^K} P(T, S_1^K|S)$$

$$= \arg\max_{T, S_1^K, T_1^K, T_1^{K'}} P(S_1^K|S)\, P(T_1^K|S_1^K, S)\, P(T_1^{K'}|T_1^K, S_1^K, S)\, P(T|T_1^{K'}, T_1^K, S_1^K, S)$$

短语划分模型

短语翻译模型

短语调序模型

目标语言模型

基于短语的翻译模型

$$T' = \arg\max_{T} P(T|S)$$

$$= \arg\max_{T, S_1^K} P(T, S_1^K|S)$$

$$= \arg\max_{T, S_1^K, T_1^K, T_1^{K'}} P(S_1^K|S)\, P(T_1^K|S_1^K, S)\, P(T_1^{K'}|T_1^K, S_1^K, S)\, P(T|T_1^{K'}, T_1^K, S_1^K, S)$$

短语划分模型

目标：将一个词序列如何划分为短语序列

方法：一般假设每一种短语划分方式都是等概率的

基于短语的翻译模型

$$T' = \arg\max_{T} P(T|S)$$

$$= \arg\max_{T, S_1^K} P(T, S_1^K|S)$$

$$= \arg\max_{T, S_1^K, T_1^K, T_1^{K'}} \underbrace{P(S_1^K|S)}\ \underbrace{P(T_1^K|S_1^K, S)}\ \underbrace{P(T_1^{K'}|T_1^K, S_1^K, S)}\ \underbrace{P(T|T_1^{K'}, T_1^K, S_1^K, S)}$$

剩下的三个核心模型：

1. **短语翻译模型：** $P(T_1^K|S_1^K, S)$

2. **短语调序模型：** $P(T_1^{K'}|T_1^K, S_1^K, S)$

3. **目标语言模型：** $P(T|T_1^{K'}, T_1^K, S_1^K, S)$

基于短语的翻译模型

短语翻译模型： $\boldsymbol{P}(T_1^K|S_1^K,\boldsymbol{S})$

1. 如何学习短语翻译规则 — 双语句对词语对齐
2. 如何估计短语翻译概率 — 短语翻译规则抽取

(澳洲 是, Australia is)

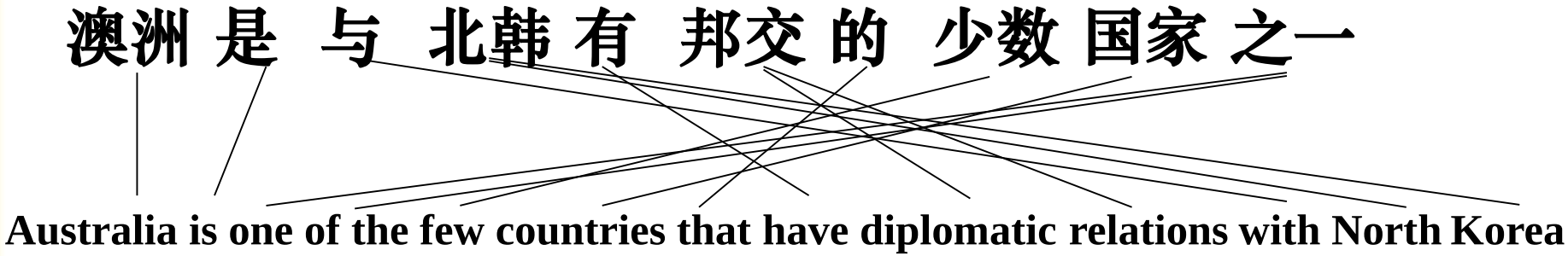
(与 北韩, with North Korea)

(有 邦交, have the diplomatic relations)

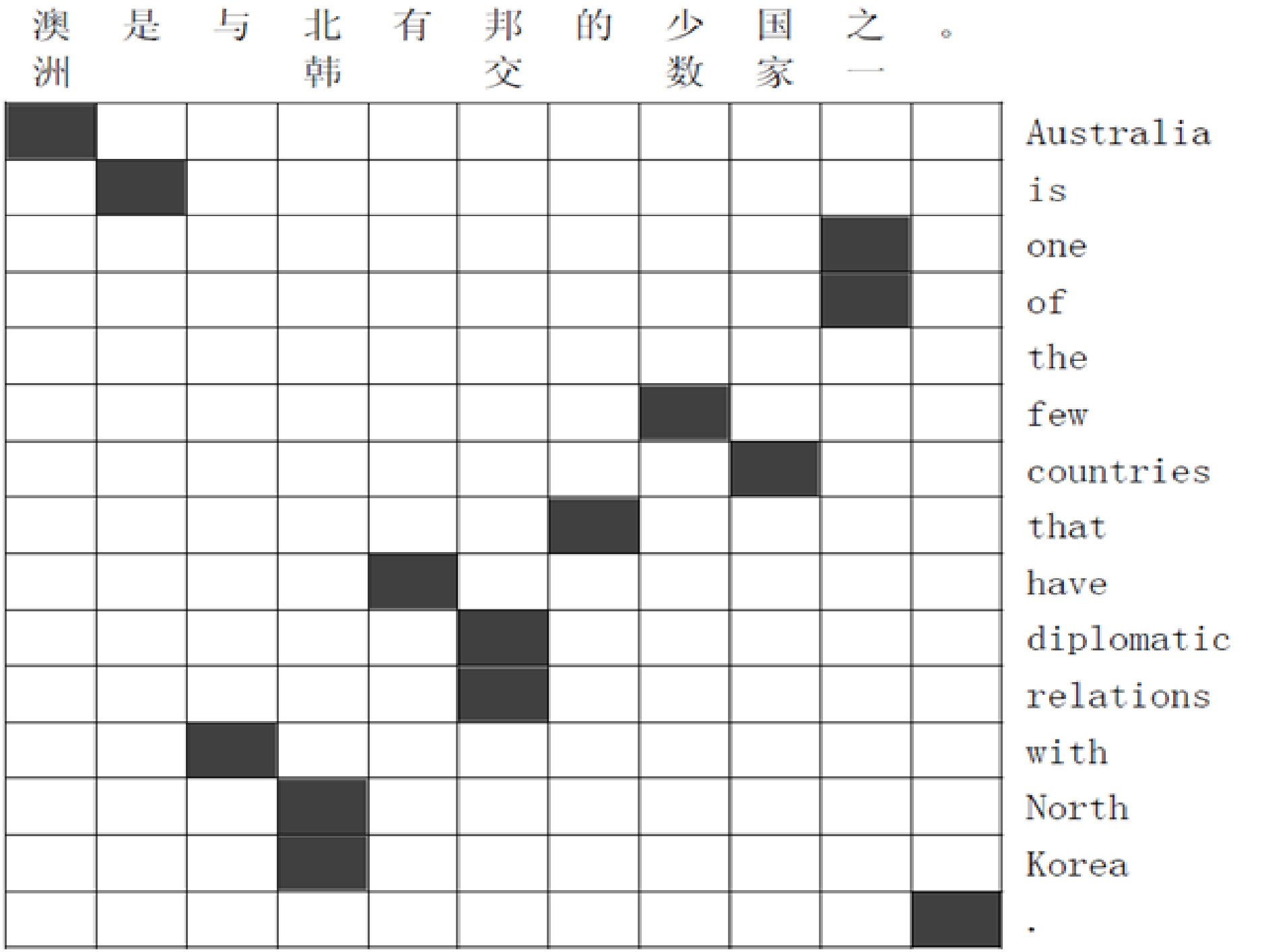
(的 少数 国家 之一, one of the few countries that)

基于短语的翻译模型

双语句对词语对齐



IBM model 1-5


澳洲
是
与
北韩
有
邦交
的
少数
国家
之一
。
Australia
is
one
of
the
few
countries
that
have
diplomatic
relations
with
North
Korea
.

基于短语的翻译模型

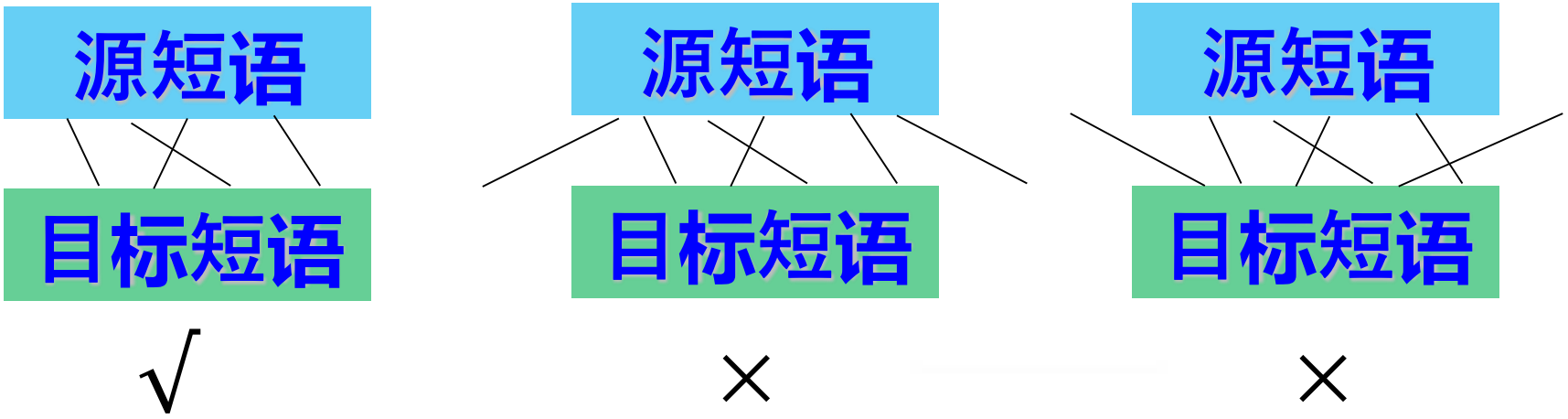
短语翻译规则抽取

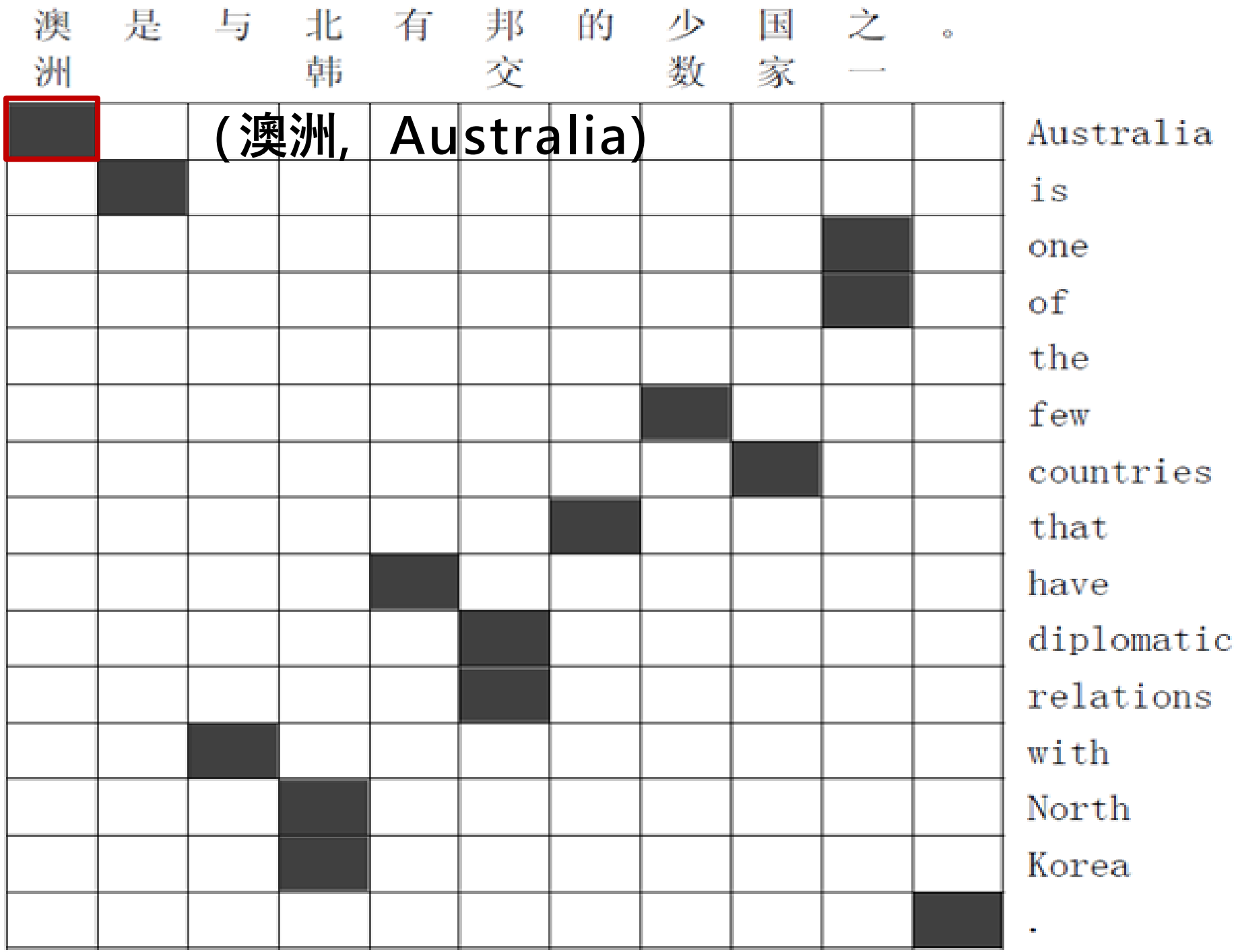
算法：对于源语言句子S中的任一短语S^j，根据词语对齐A找到目标语言句子T中的对齐片段$T_{i'}^{j'}$，若S^j与$T_{i'}^{j'}$满足对齐一致性，则$\left(S^j, T_{i'}^{j'}\right)$为一条短语翻译规则。

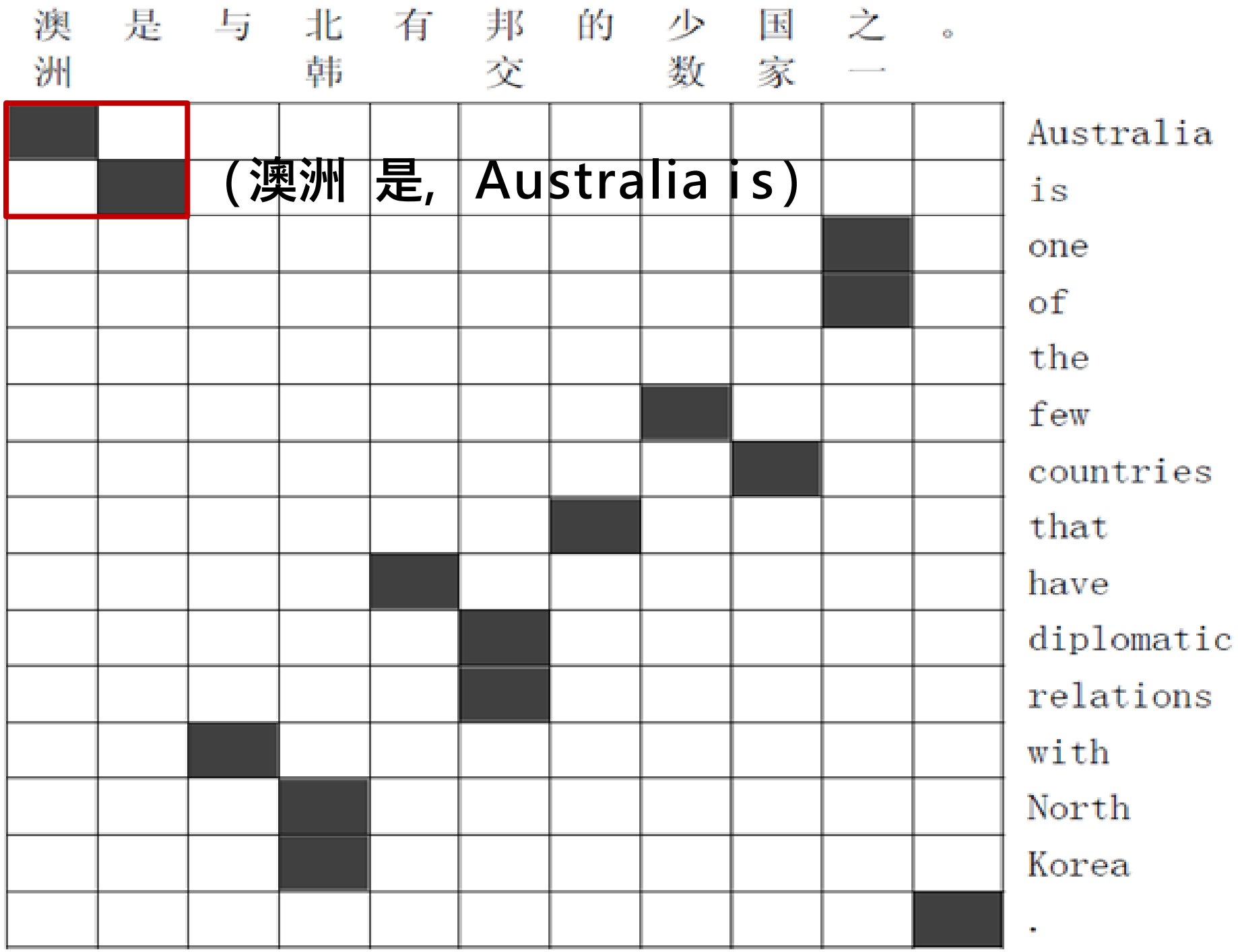
基于短语的翻译模型

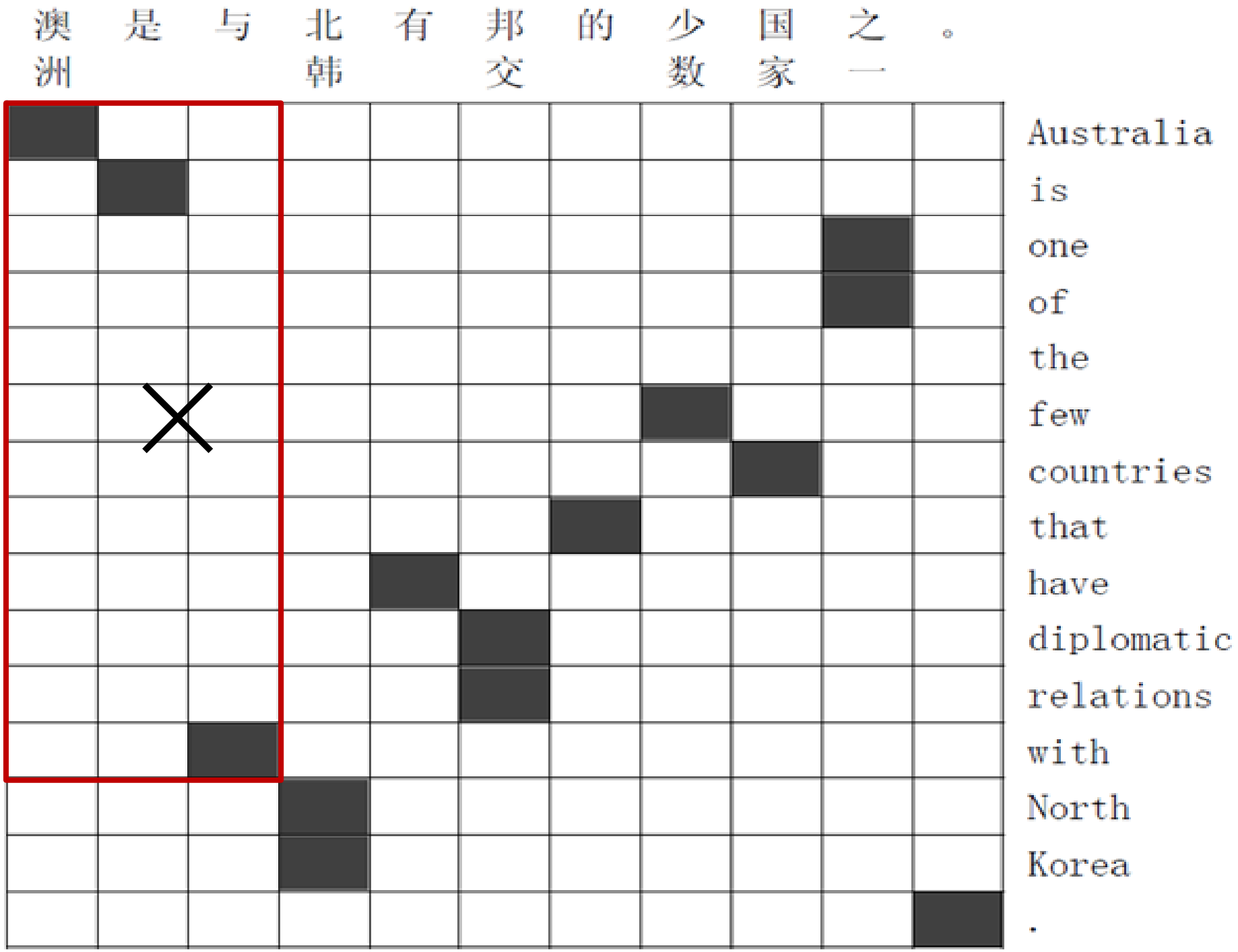
短语翻译规则抽取

对齐一致性： S^j**中每个词**S_k**，若**$(k,k')\in A$**，则**$i'\le k'\le j'$**，** $T_{i'}^{j'}$ **中每个词**$T_{t'}$**，若**$(t,t')\in A$**，则**$i\le t\le j$**。**

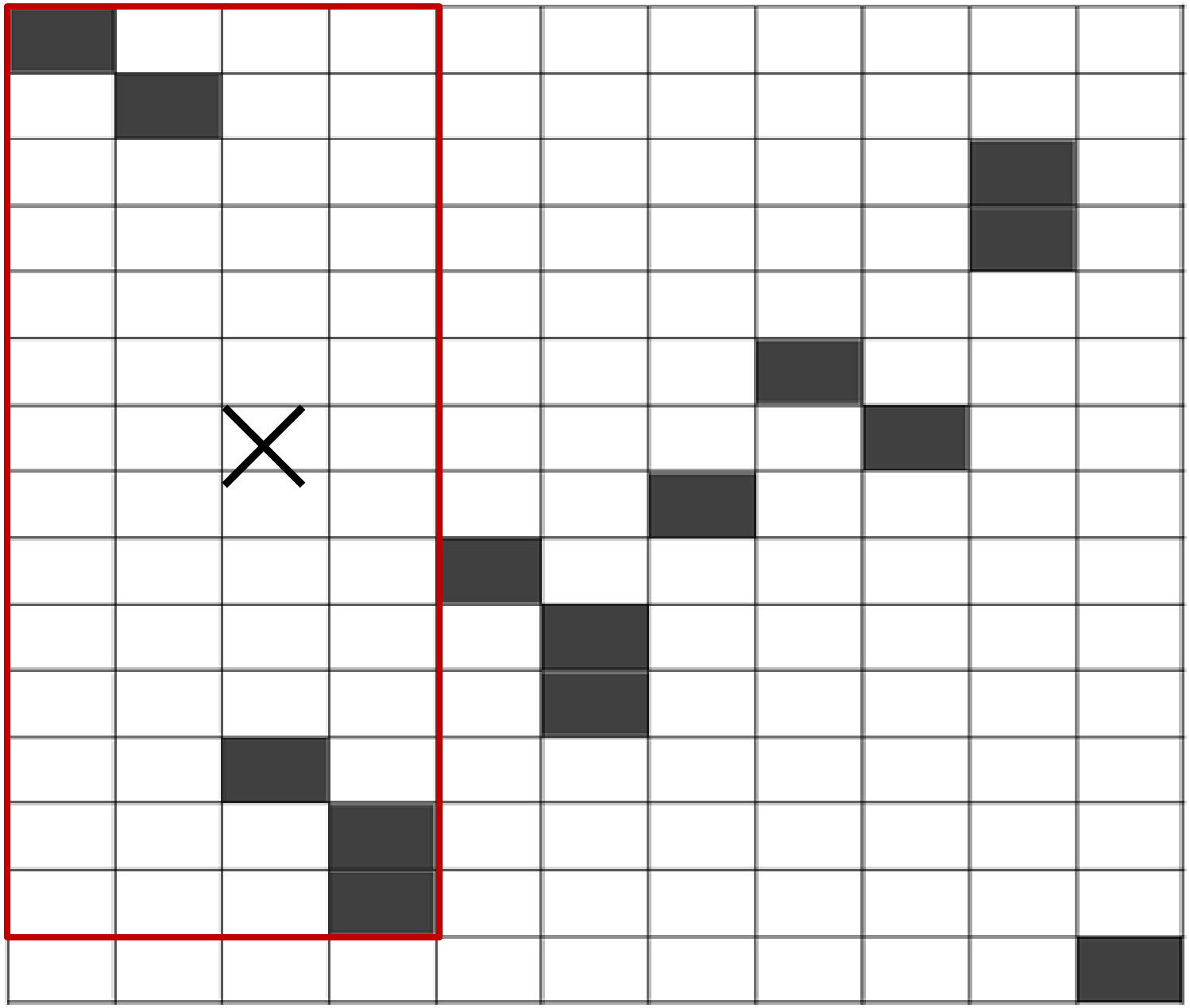



澳洲 是 与 北韩 有 邦交 的 少数 国家 之一 。
(澳洲, Australia)
Australia
is
one
of
the
few
countries
that
have
diplomatic
relations
with
North
Korea
.


澳洲 是 与 北韩 有 邦交 的 少数 国家 之一 。
（澳洲 是, Australia is）
Australia
is
one
of
the
few
countries
that
have
diplomatic
relations
with
North
Korea
.

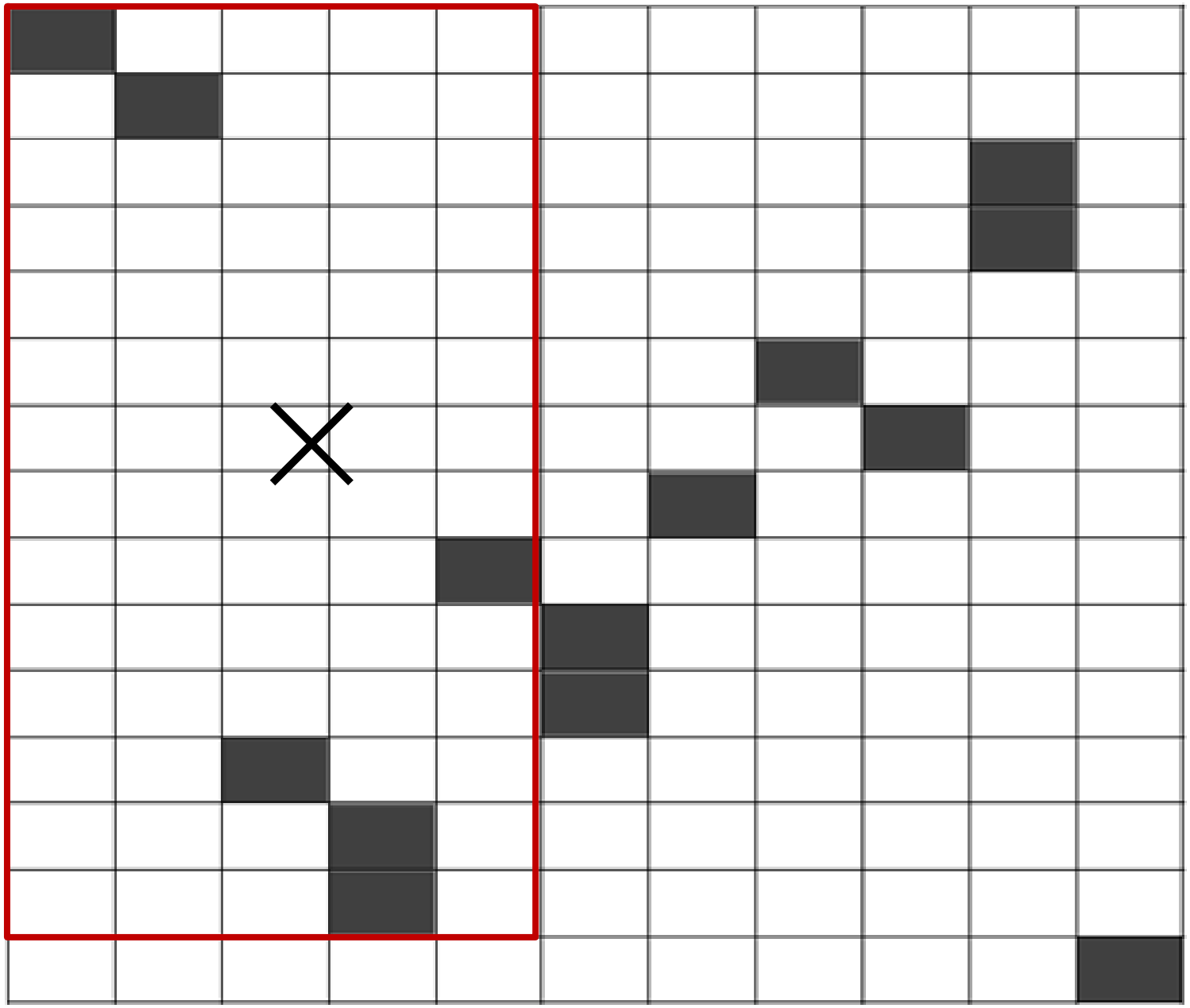

澳洲
是
与
北韩
有
邦交
的
少数
国家
之一
。
Australia
is
one
of
the
few
countries
that
have
diplomatic
relations
with
North
Korea
.

澳洲 是 与 北韩 有 邦交 的 少数 国家 之一 。

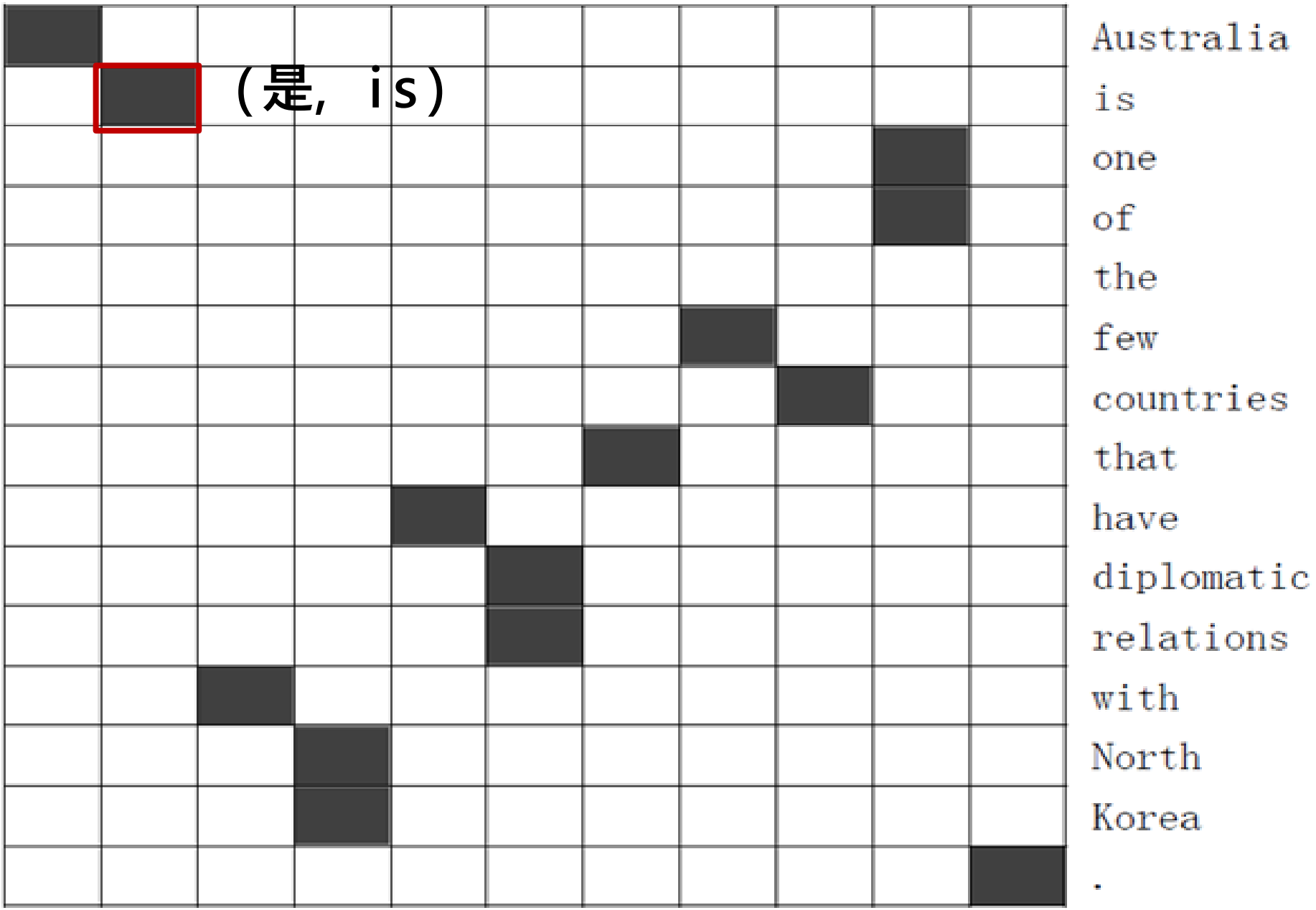


Australia is one of the few countries that have diplomatic relations with North Korea .

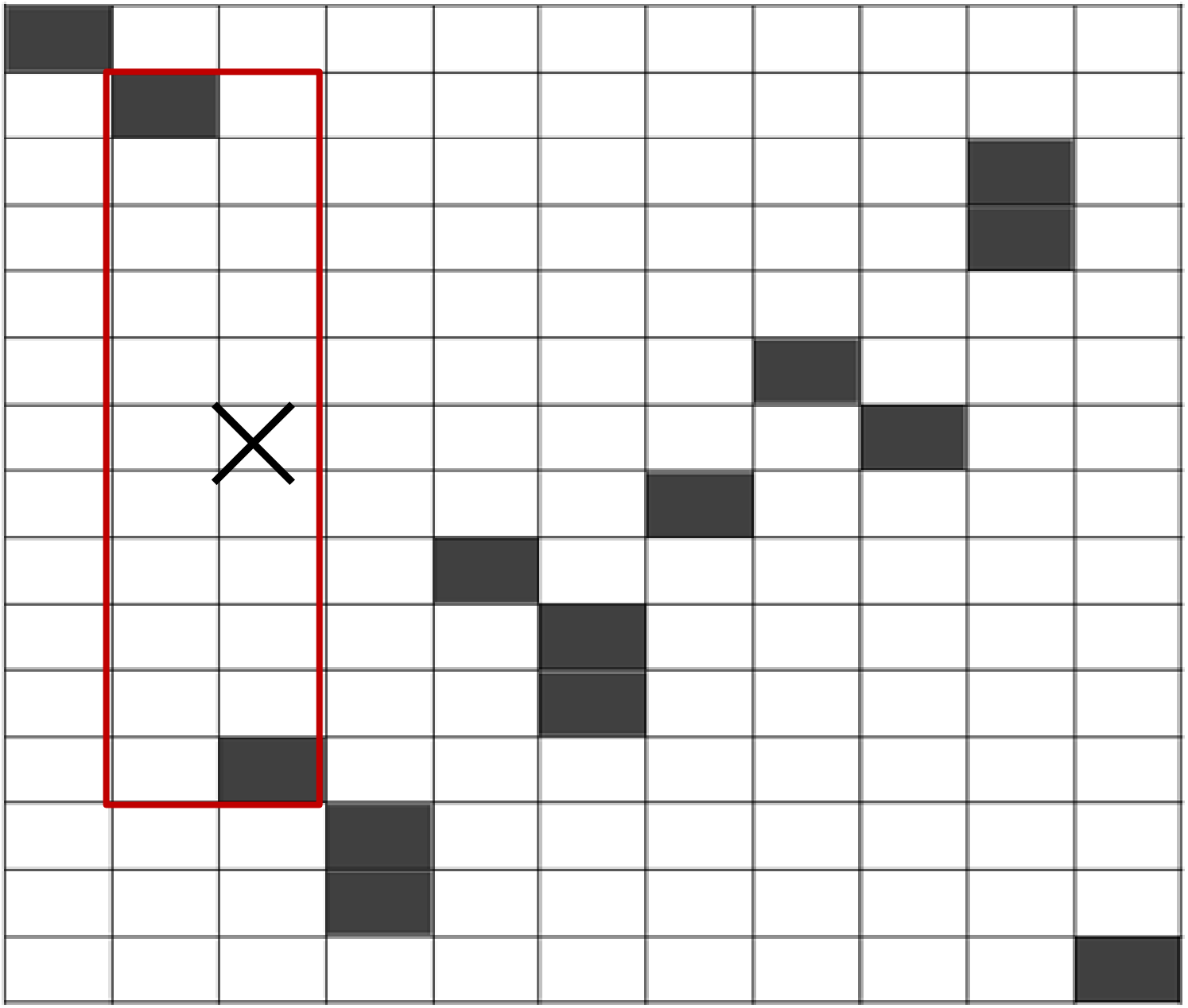
澳洲 是 与 北韩 有 邦交 的 少数 国家 之一 。



Australia is one of the few countries that have diplomatic relations with North Korea .

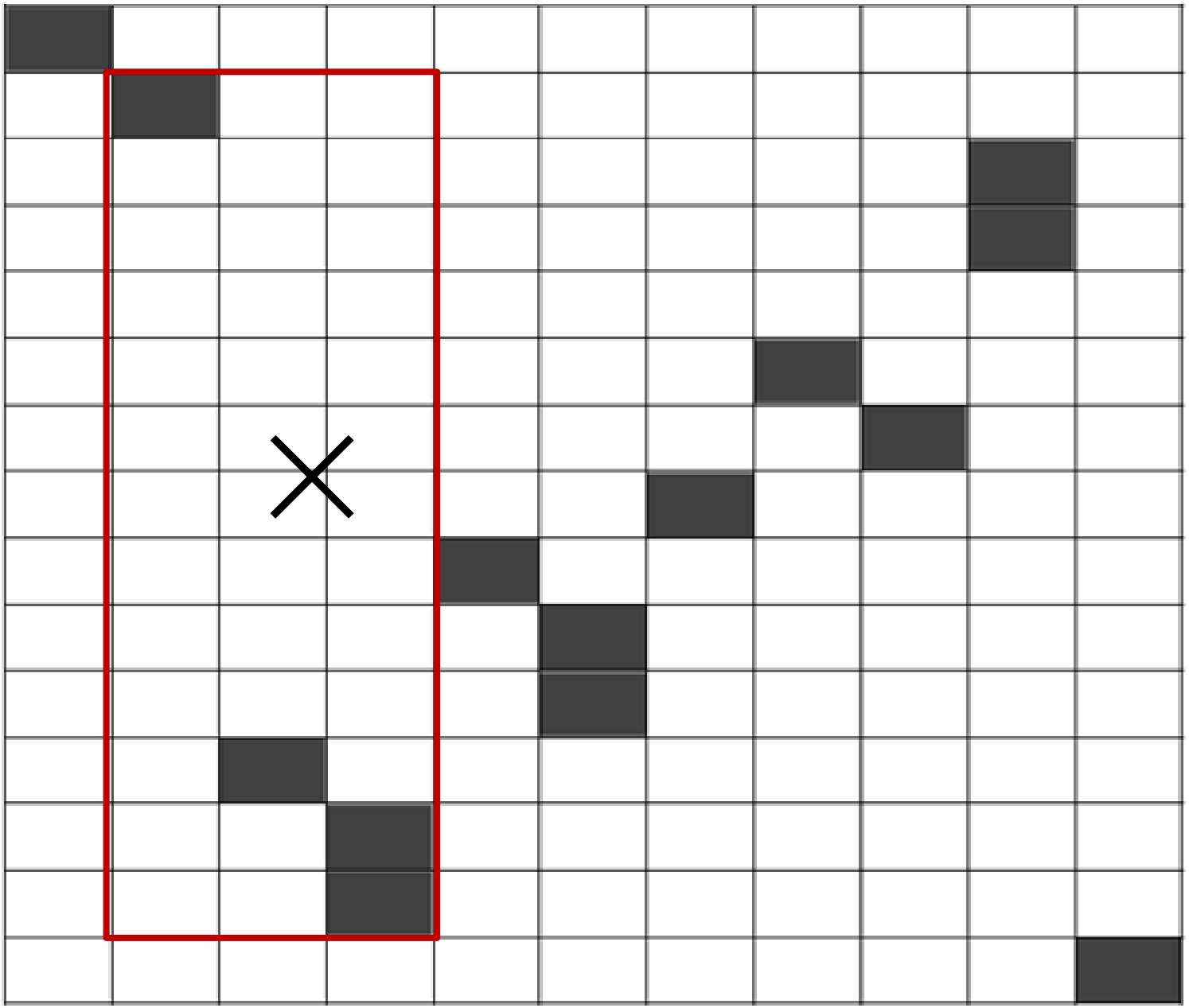

澳洲
是
与
北韩
有
邦交
的
少数
国家
之一
。
（是，is）
Australia
is
one
of
the
few
countries
that
have
diplomatic
relations
with
North
Korea
.

澳洲 是 与 北韩 有 邦交 的 少数 国家 之一 。



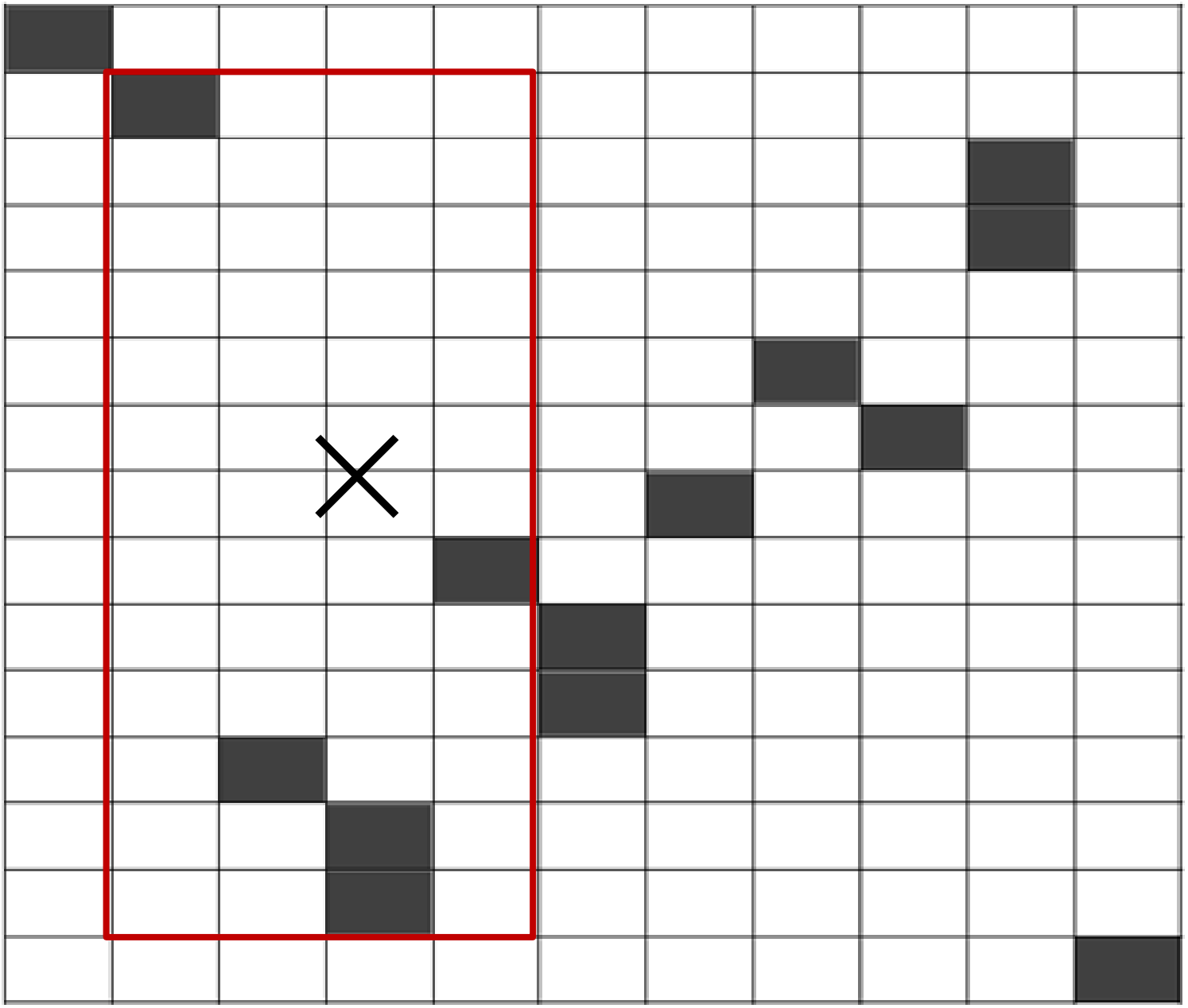
Australia is one of the few countries that have diplomatic relations with North Korea .

澳洲 是 与 北韩 有 邦交 的 少数 国家 之一 。



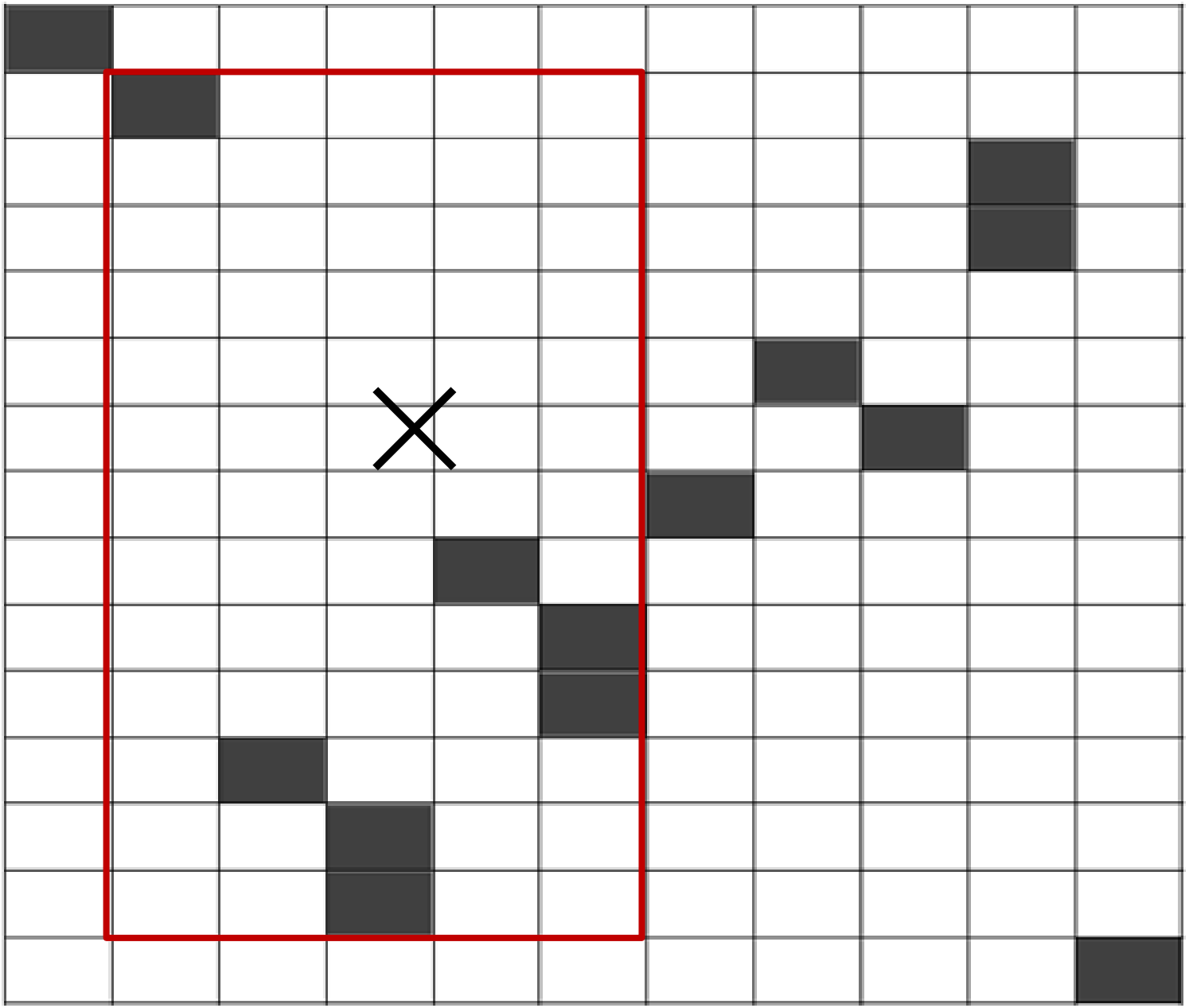
Australia is one of the few countries that have diplomatic relations with North Korea .

澳洲 是 与 北韩 有 邦交 的 少数 国家 之一 。

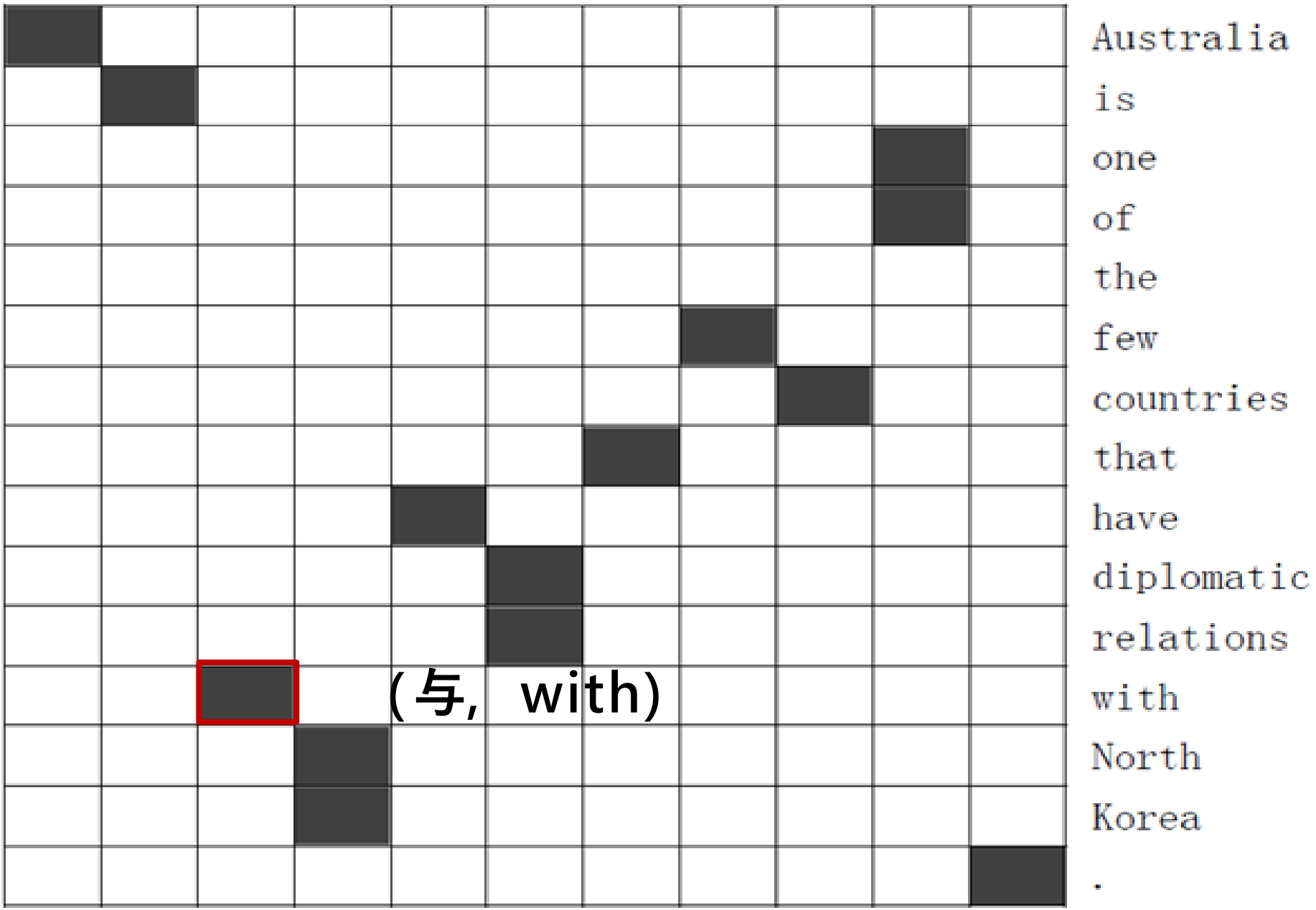


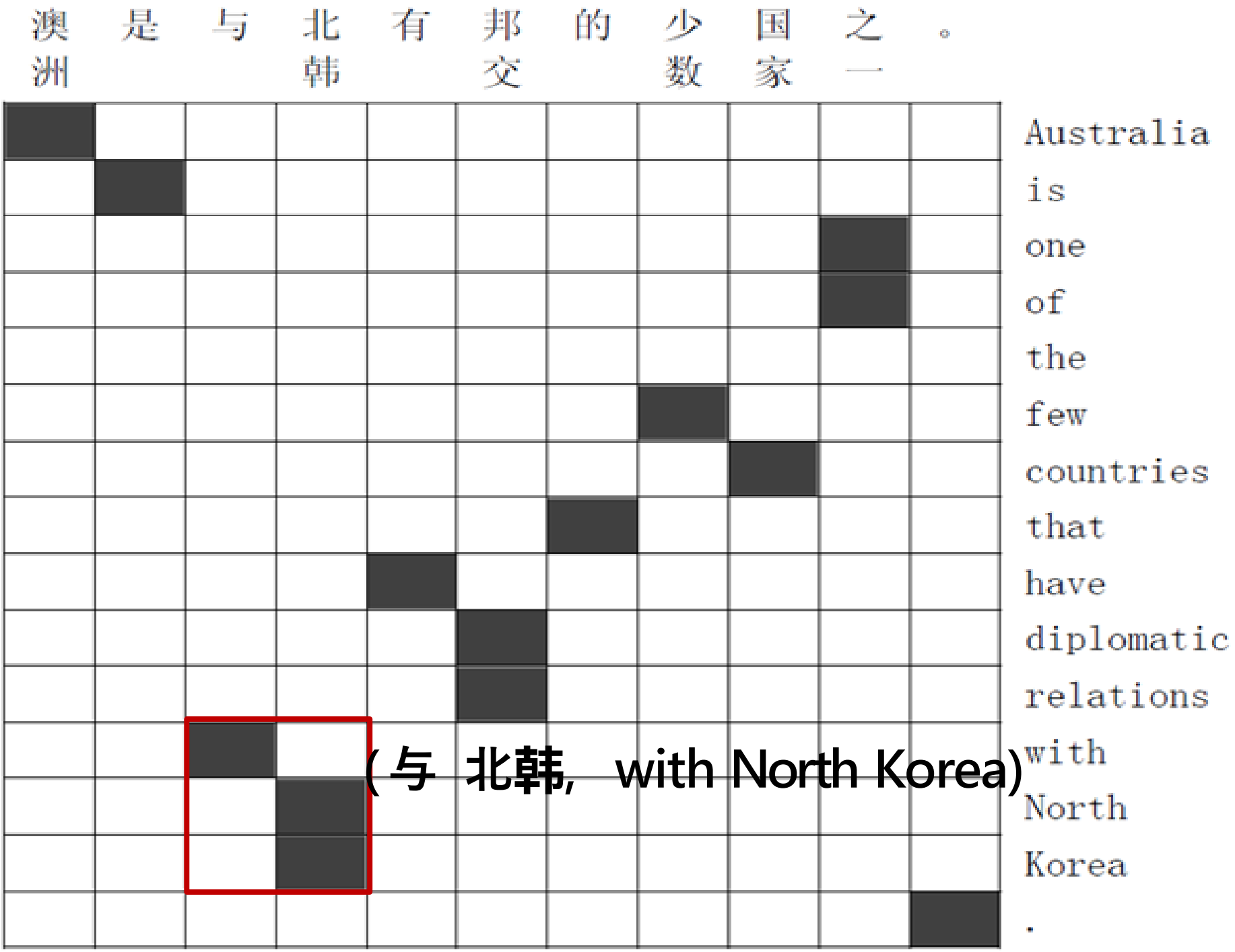
Australia is one of the few countries that have diplomatic relations with North Korea .

澳洲 是 与 北韩 有 邦交 的 少数 国家 之一 。

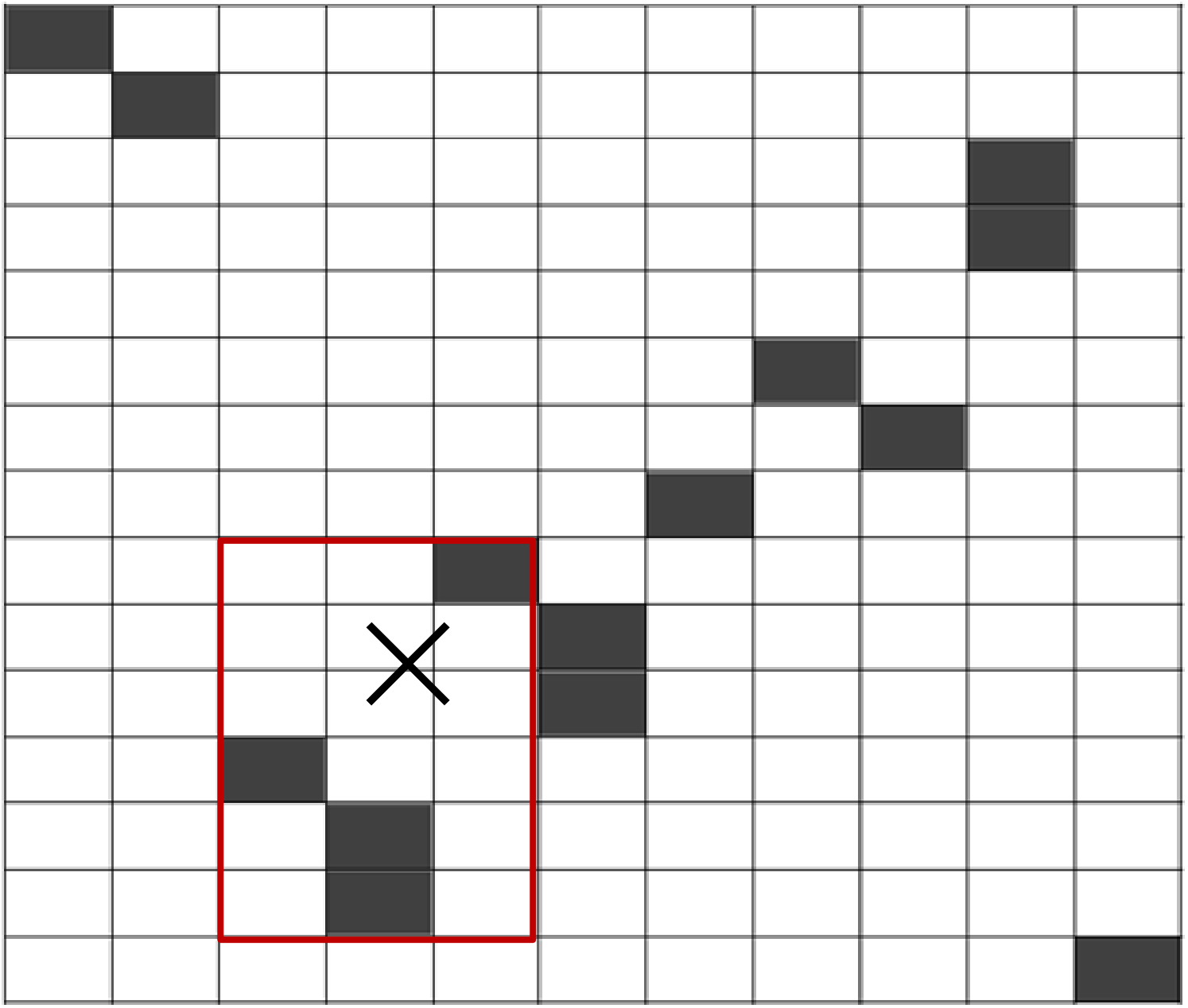


Australia is one of the few countries that have diplomatic relations with North Korea .


澳洲
是
与
北韩
有
邦交
的
少数
国家
之一
。
Australia
is
one
of
the
few
countries
that
have
diplomatic
relations
with
North
Korea
.
(与，with)

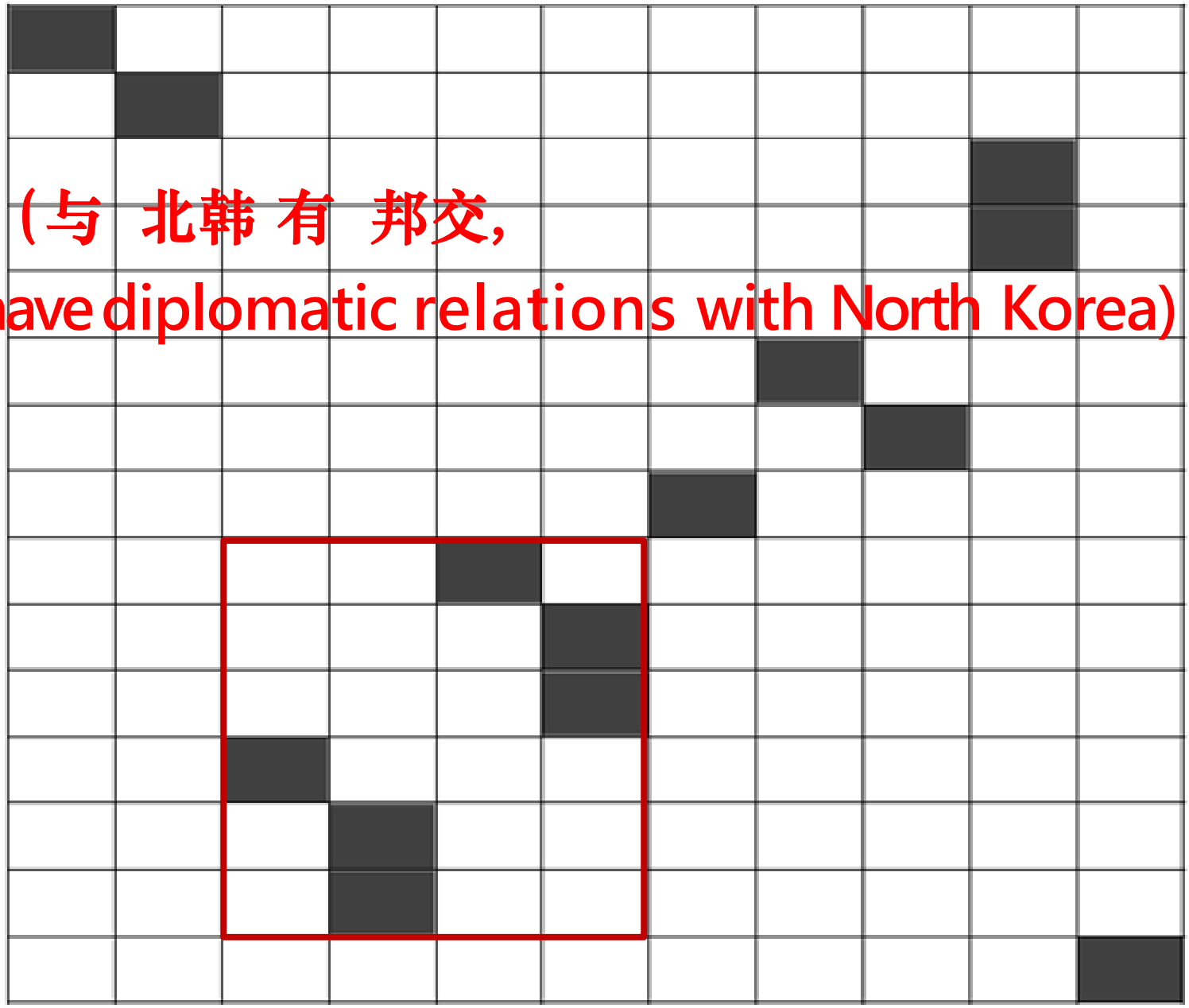

澳洲 是 与 北韩 有 邦交 的 少数 国家 之一 。
Australia
is
one
of
the
few
countries
that
have
diplomatic
relations
with
North
Korea
.
(与 北韩, with North Korea)

澳洲 是 与 北韩 有 邦交 的 少数 国家 之一 。



Australia is one of the few countries that have diplomatic relations with North Korea .

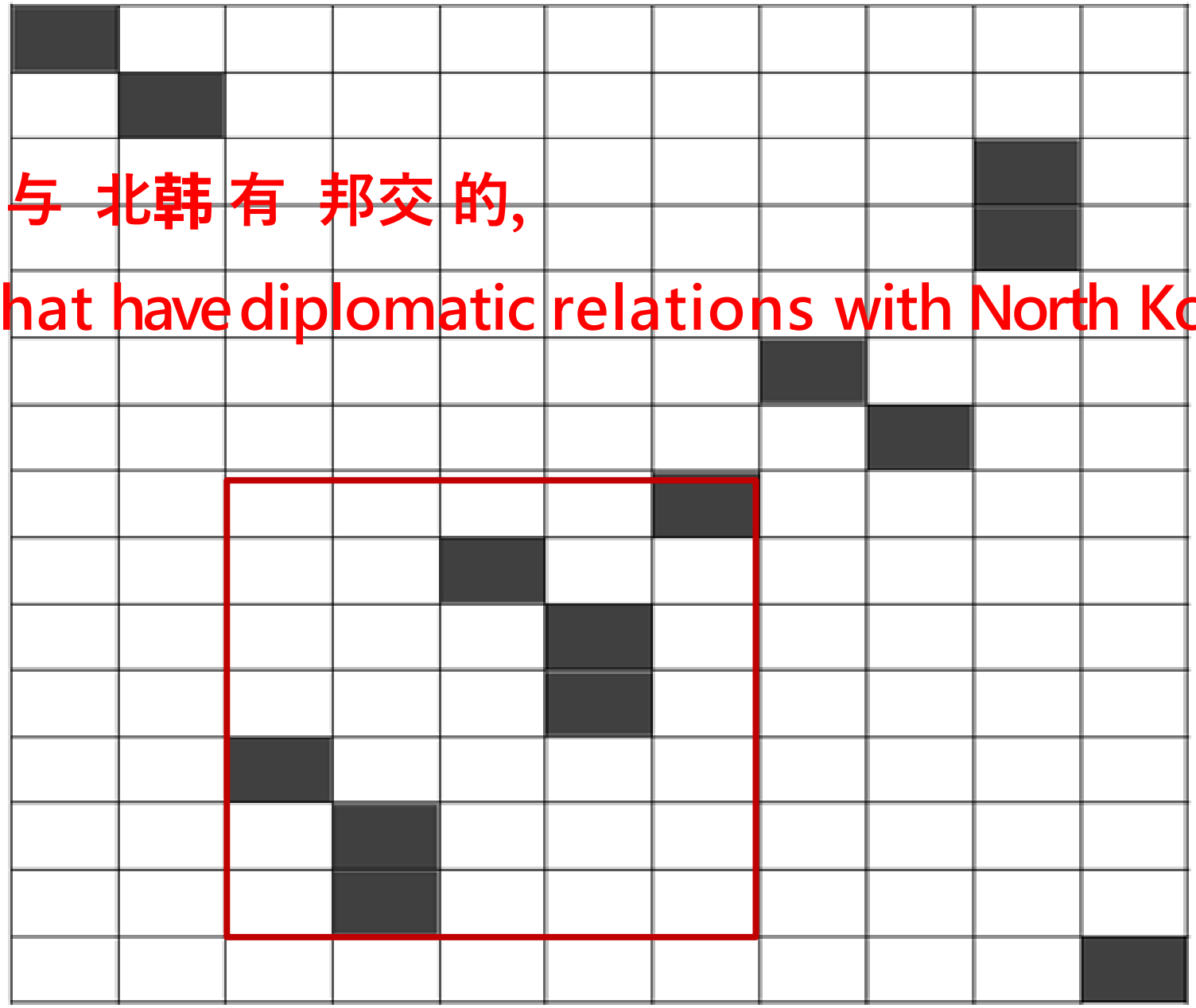
澳洲 是 与 北韩 有 邦交 的 少数 国家 之一 。



(与 北韩 有 邦交,
have diplomatic relations with North Korea)

Australia is one of the few countries that have diplomatic relations with North Korea .

澳洲 是 与 北韩 有 邦交 的 少数 国家 之一 。



Australia is one of the few countries that have diplomatic relations with North Korea .

(与 北韩 有 邦交 的,

that have diplomatic relations with North Korea)

基于短语的翻译模型

短语翻译模型： $\boldsymbol{P}(T_1^K|S_1^K, \boldsymbol{S})$

1. 如何学习短语翻译规则
   - 双语句对词语对齐
   - 短语翻译规则抽取
2. 如何估计短语翻译概率

(澳洲 是, Australia is)

(与 北韩, with North Korea)

(有 邦交, have the diplomatic relations)

(的 少数 国家 之一, one of the few countries that)

基于短语的翻译模型

短语翻译概率估计：4个翻译概率 （最大似然）

1. 正向、逆向短语翻译概率 $p(t|s), p(s|t)$

2. 正向、逆向词汇化翻译概率 $p_{lex}(t|s), p_{lex}(s|t)$

（与 北韩， with North Korea)

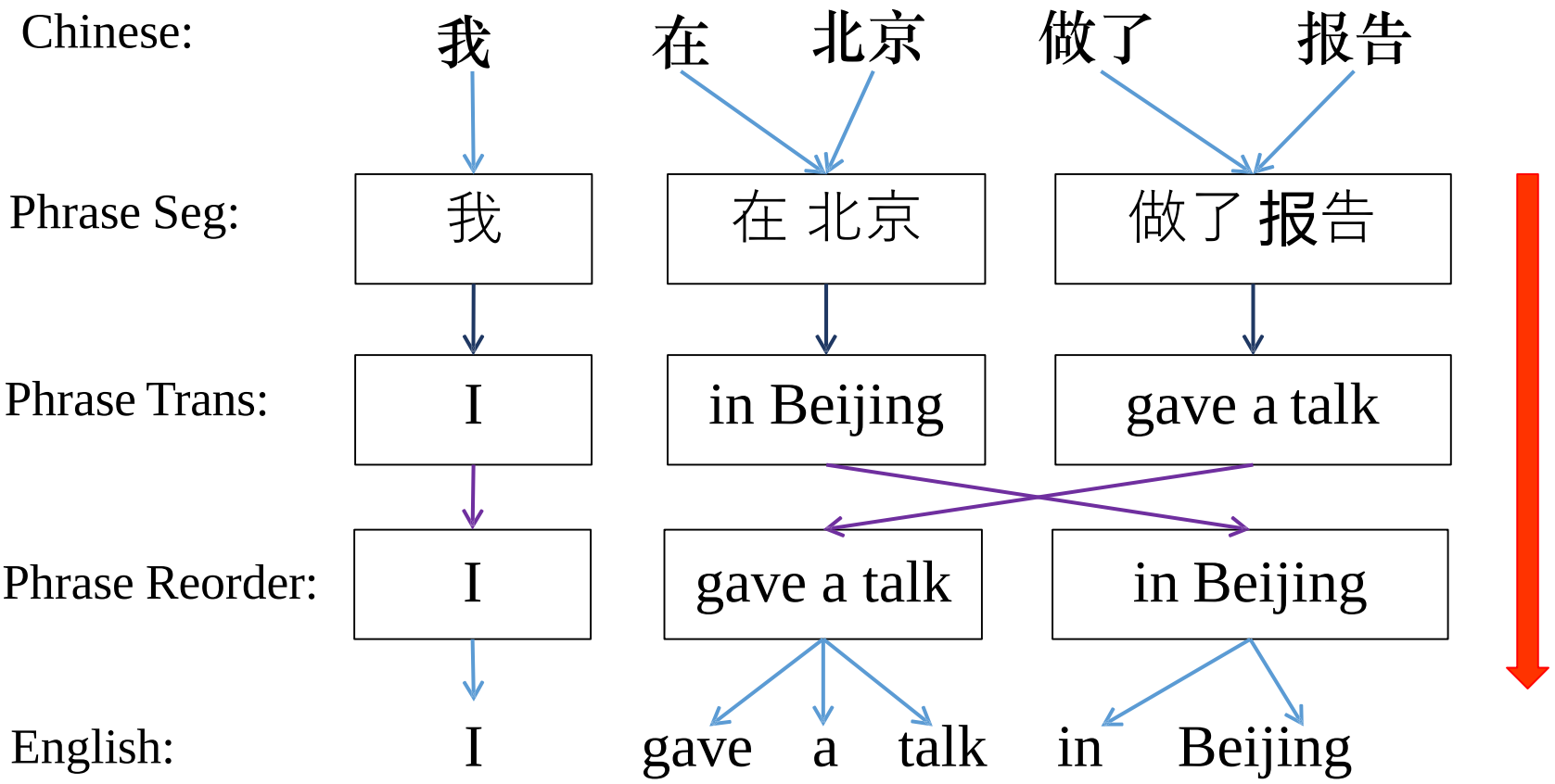
$p(with|\text{与}) = 0.4$

$p(North|\text{北韩}) = 0.1$

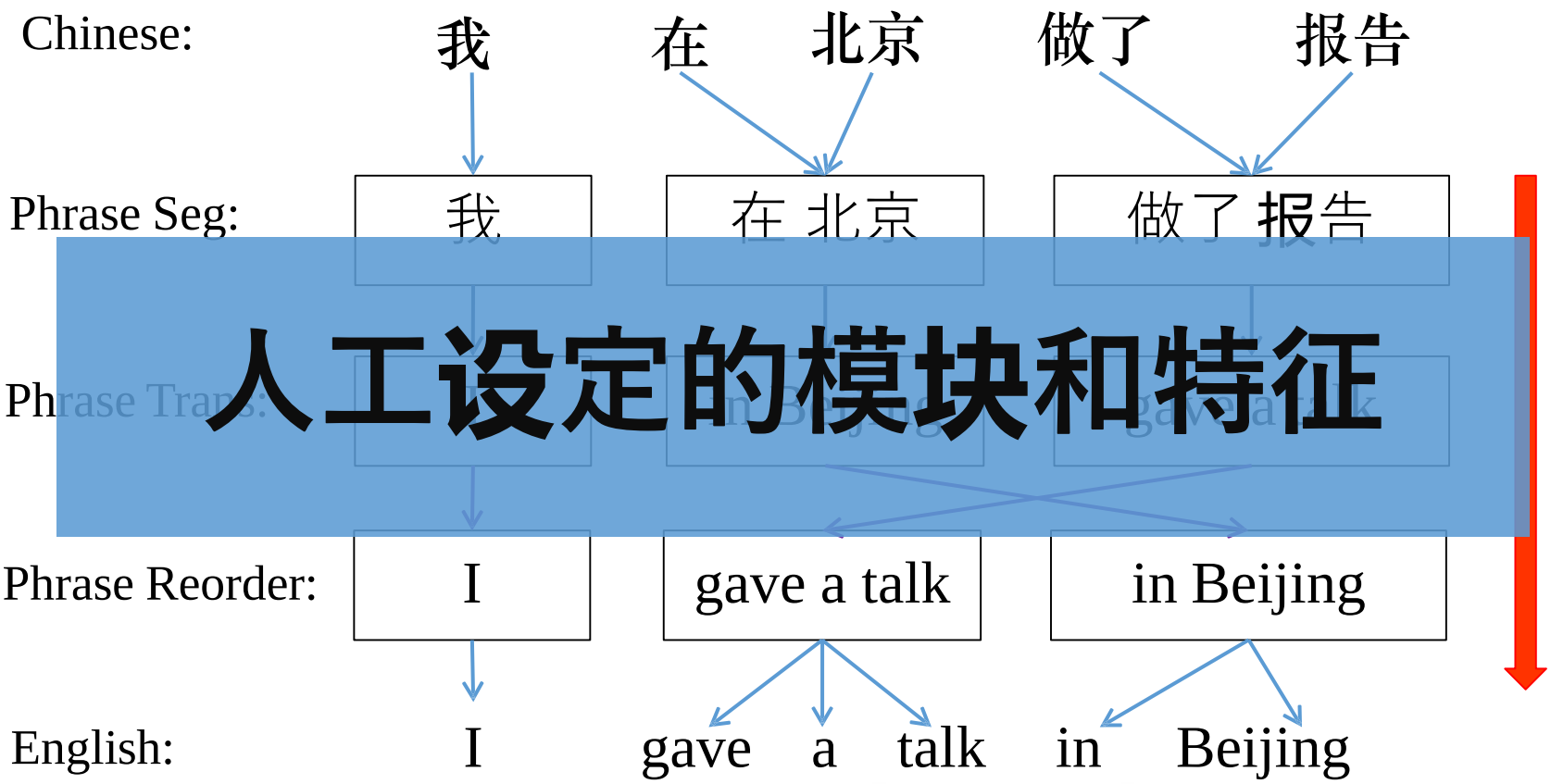
$p(Korea|\text{北韩}) = 0.5$

$p_{lex}(t|s) = 0.4 \times 0.1 \times 0.5 = 0.02$

统计机器翻译

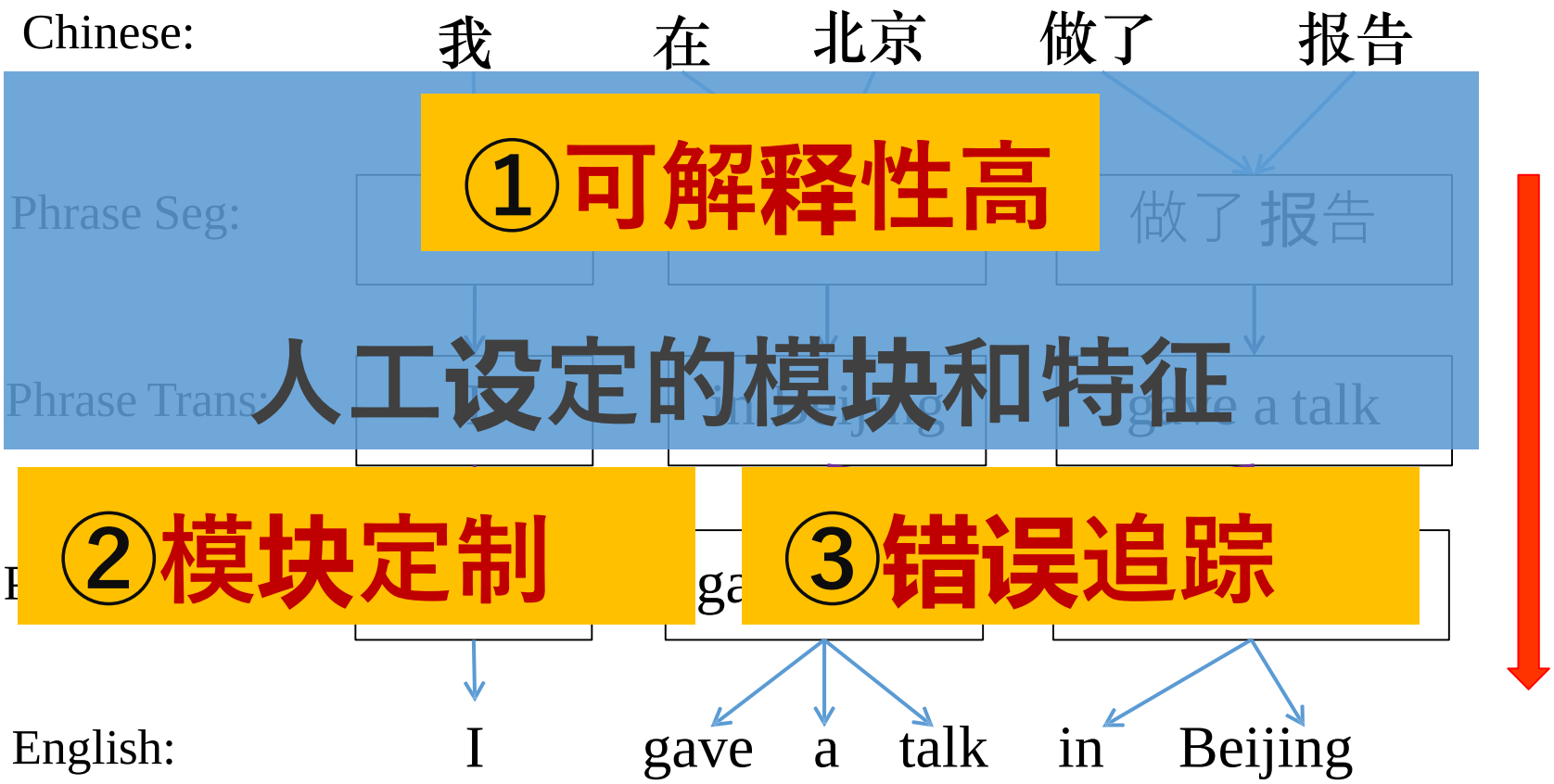

Chinese:
我
在
北京
做了
报告
Phrase Seg:
我
在 北京
做了 报告
Phrase Trans:
I
in Beijing
gave a talk
Phrase Reorder:
I
gave a talk
in Beijing
English:
I
gave
a
talk
in
Beijing

统计机器翻译

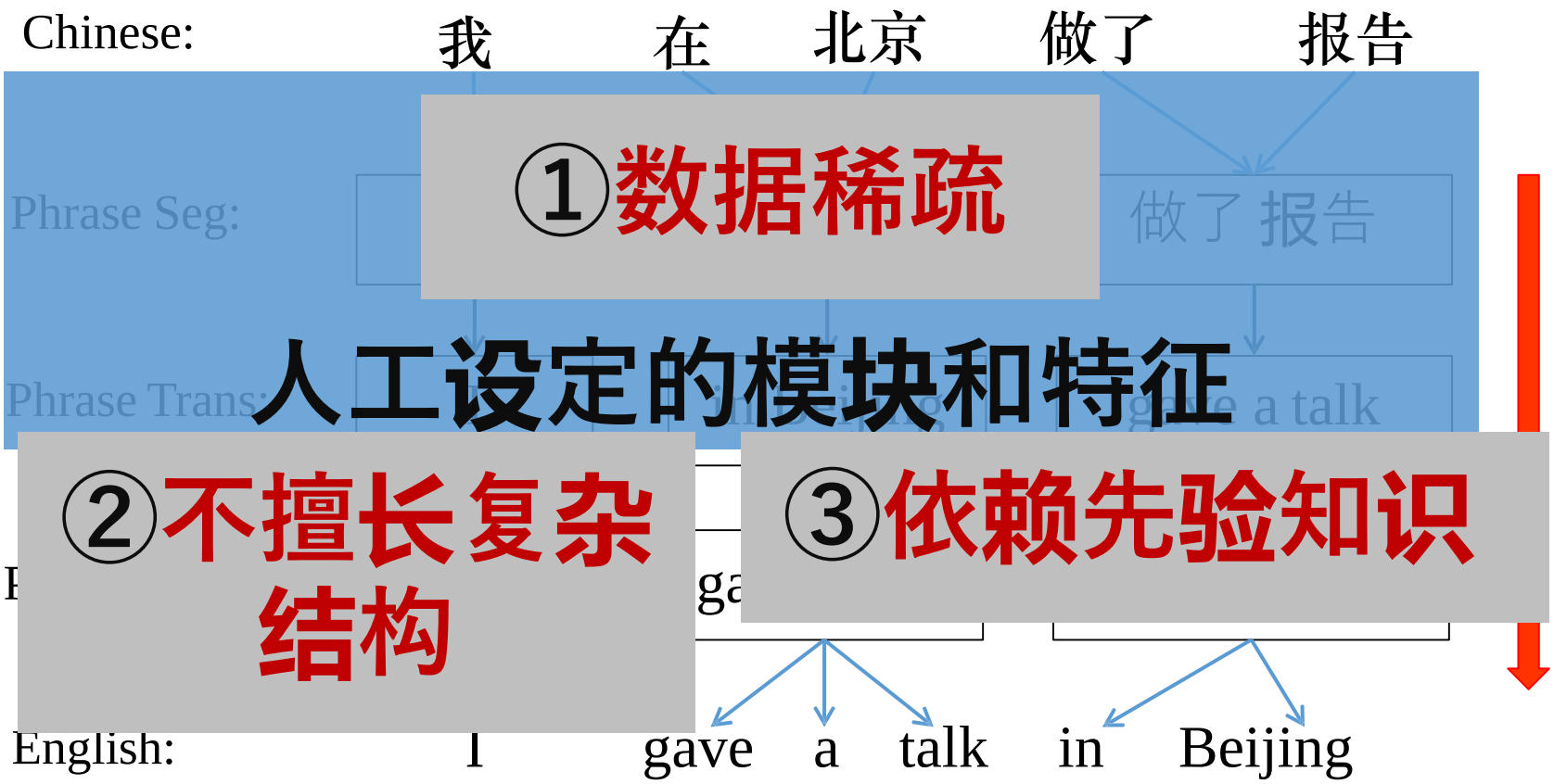


人工设定的模块和特征

统计机器翻译


Chinese:
我
在
北京
做了
报告
Phrase Seg:
做了 报告
Phrase Trans:
gave a talk
①可解释性高
人工设定的模块和特征
②模块定制
③错误追踪
English:
I
gave
a
talk
in
Beijing

统计机器翻译



①数据稀疏

人工设定的模块和特征

②不擅长复杂结构

③依赖先验知识

统计机器翻译→神经机器翻译

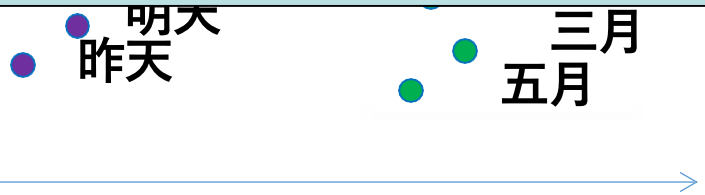
离散符号表示方法 ⇨ 连续分布式表示方法

讲座 ⊗ 报告 = 0

讲座 报告

$$\begin{bmatrix}0.48\\0.46\\0.26\end{bmatrix} \otimes \begin{bmatrix}0.42\\0.51\\0.21\end{bmatrix} \approx 1$$

分布式的语义表示是统计机器翻译到神经机器翻译的核心

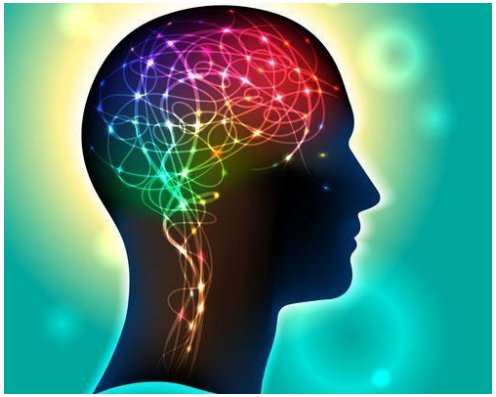


低维、稠密的连续实数空间

神经机器翻译

Chinese: 我 在 北京 做了 报告

编码网络



分布式语义表示

解码网络

English: I gave a talk in Beijing

神经机器翻译

Chinese: 我 在 北京 做了 报告

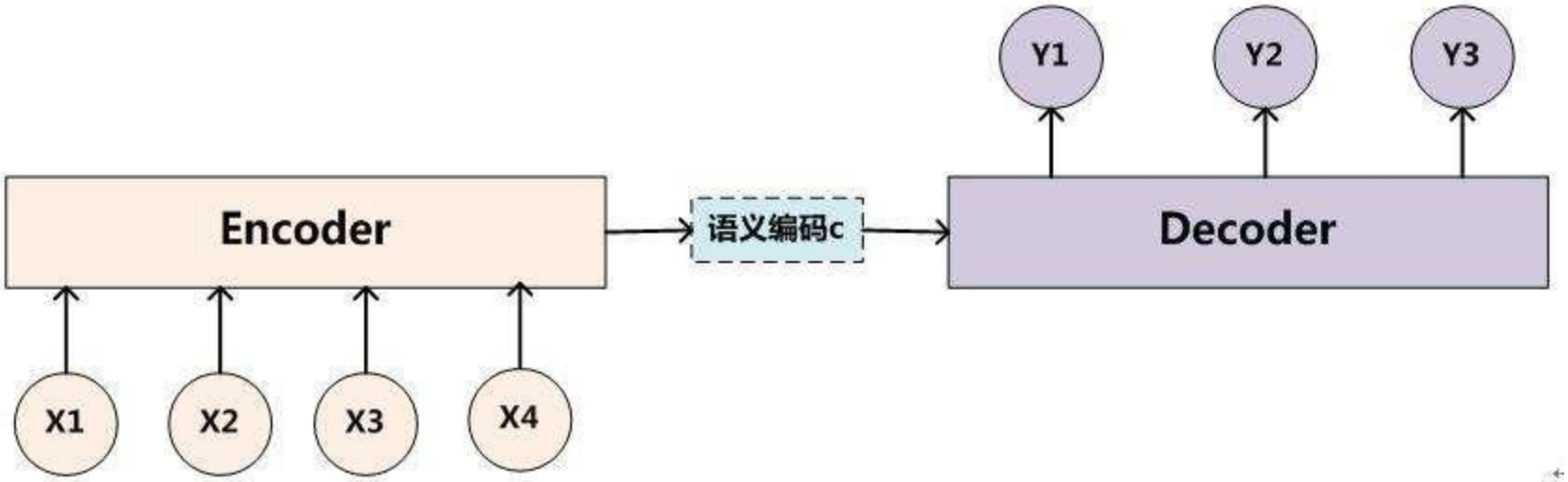
编码网络

仅需要两个神经网络

解码网络

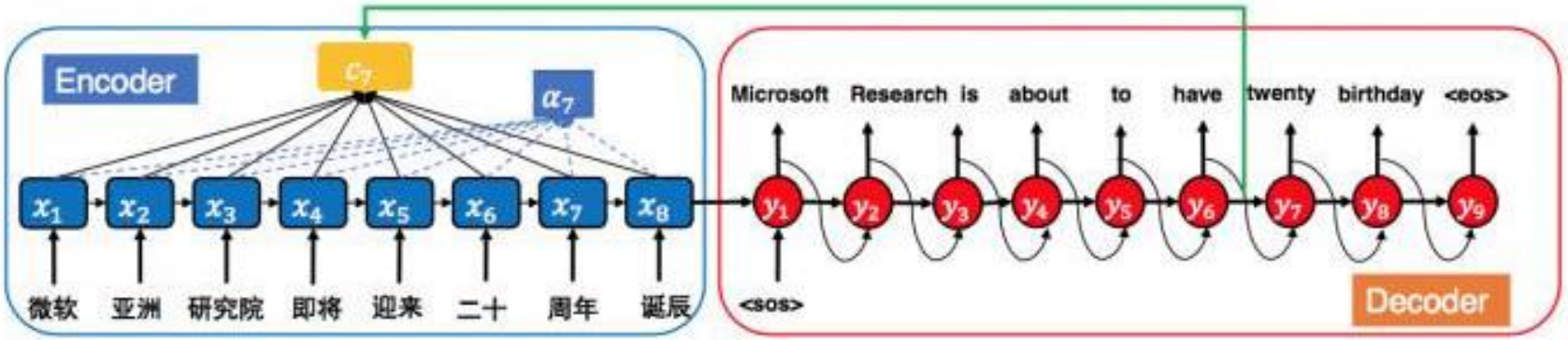
English: I gave a talk in Beijing

神经机器翻译



统计机器翻译→神经机器翻译

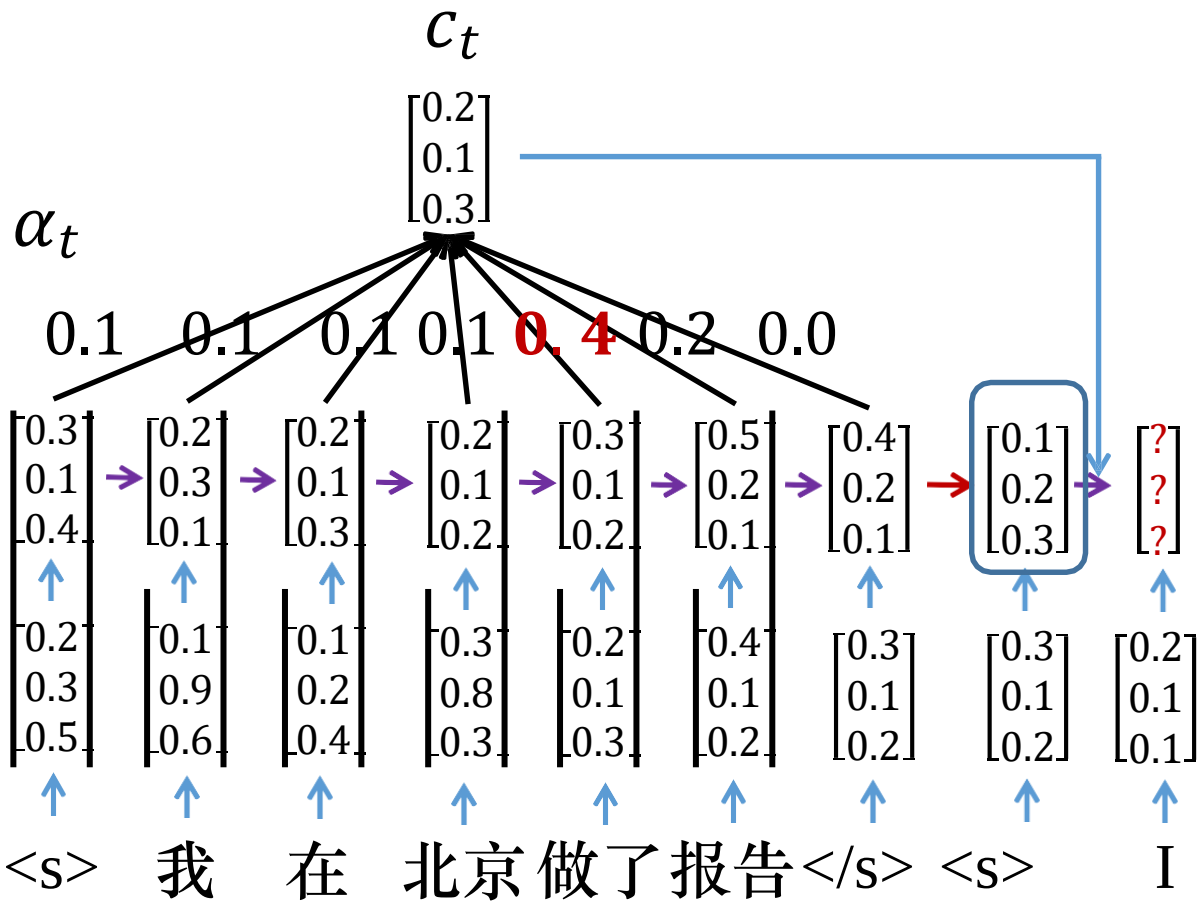
Attention 注意力机制



- $c_j = \sum_{i=1}^{T_x} \alpha_{ji} h_i$
- $\alpha_{ji} = \frac{\exp(e_{ji})}{\sum_{k=1}^{T_x} \exp(e_{jk})}$
- $e_{ji} = A(s_{j-1}, h_i)$

$s_j = f(y_{j-1}, s_{j-1}; c_j)$: LSTM/GRU

神经机器翻译-注意机制



译文评估方法

- **BLEU评价方法 [Papineni, 2002]**
  - **BiLingual Evaluation Understudy, IBM**

- **基本思想**：将机器翻译产生的候选译文与人翻译的多个参考译文相比较，越接近，候选译文的正确率越高

- **实现方法**：统计同时出现在系统译文和参考译文中的*n*元词的个数，最后把匹配到的*n*元词的数目除以系统译文的n元词数目，得到评测结果。

Thank you!

权小军 中山大学数据科学与计算机学院