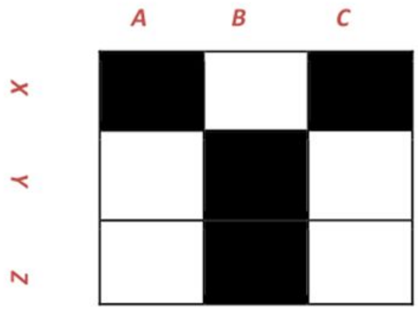
Statistical Machine Translation

Lab Exercise

6: Phrase-based Model

Please use Java as your programming language for this lab
Refer to the lecture slide (Week 7) for extra information

1- Given the word alignment between the source sentence "X Y Z" and the target sentence "A B B" as follows:

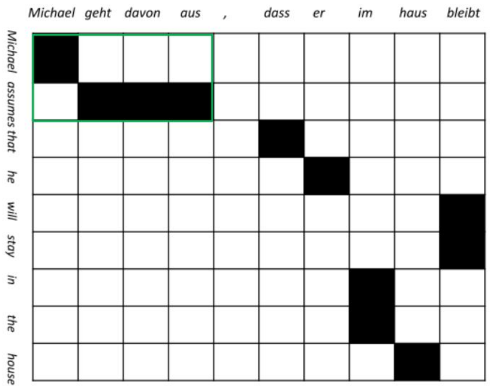


Please **manually** extract all phrase pairs that are consistent with the word alignment.

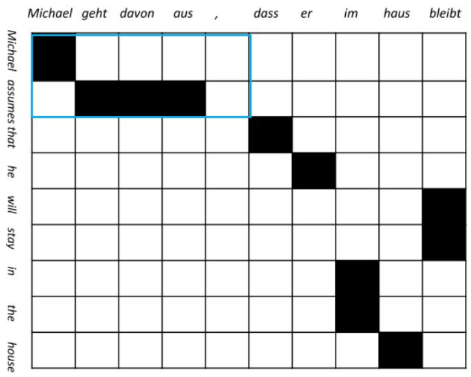
2- Given a French (source) and an English sentences (target) as well as their word alignment links as follows:

Source	**Target**	**Alignment**
oh , c' est quoi ton problème ?	oh , what is your deal ?	0-0 1-1 2-2 4-2 3-3 5-4 6-5 7-6

2.1 Please follow the consistency principle of phrase extraction to write a program to extract and output all possible phrase pairs. **To make the algorithm simple, we only consider the aligned situations like (a) without considering the unaligned situations like (b)**.

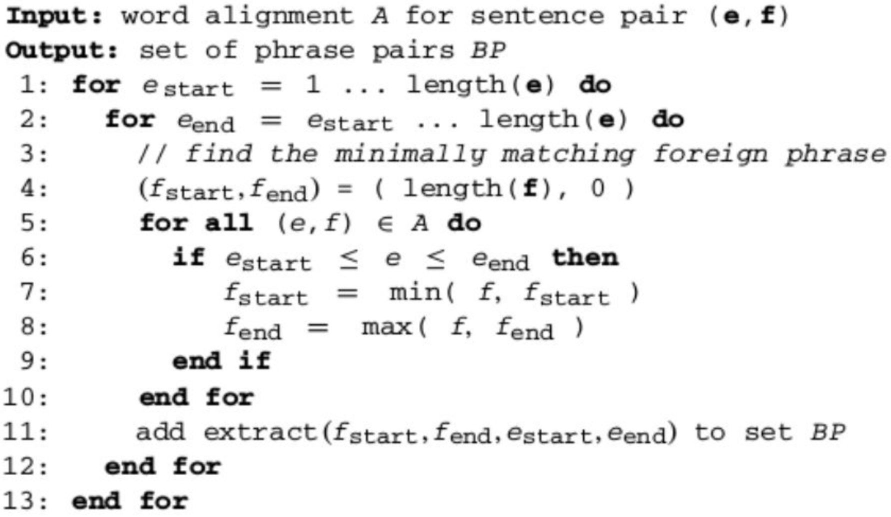


(a)



(b)

Hint: you can design your program according to the pseudo as follows:



```
Input: word alignment $A$ for sentence pair ($\mathbf{e}$, $\mathbf{f}$)
Output: set of phrase pairs $BP$
for $e_{start}$ = 1 ... length($\mathbf{e}$) do
  for $e_{end}$ = $e_{start}$ ... length($\mathbf{e}$) do
    // find the minimally matching foreign phrase
    ($f_{start}$, $f_{end}$) = ( length($\mathbf{f}$), 0 )
    for all $(e, f) \in A$ do
      if $e_{start} \leq e \leq e_{end}$ then
        $f_{start}$ = min( $f$, $f_{start}$ )
        $f_{end}$ = max( $f$, $f_{end}$ )
      end if
    end for
    add extract($f_{start}$, $f_{end}$, $e_{start}$, $e_{end}$) to set $BP$
  end for
end for
```

Output format:

c’ ||| what

2.2 Please follow the phrase probability estimation method (relative frequency) to estimate the probabilities of extracted phrases in 2.1.

Hint: the relative frequency is calculate as follows:

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

where $\bar{f}$ is source language and $\bar{e}$ is target language.

Output format:

c’ ||| what ||| 1.0

3- Following Question 1&2, given a source training file (train.fr) and a target training file (train.en) as well as their word alignment links file (align.fr-en), please write a program extract all possible parallel phrases following the consistency principle of phrase extraction, and then estimate the probabilities for all phrase pairs. The generated results are exported into a file (phrase-table.txt).

Format of output file:

oh ||| oh ||| 1.0

c’ ||| what ||| 1.0