

Übung 6

Phrasenextraktion

Extrahieren Sie für die unten gegebene Alignierungsmatrix alle Phrasenpaare, die mit der Wortalignierung konsistent sind.

- Wie viele verschiedene zusammenhängende englische und deutsche Wortfolgen enthält dieses Beispiel? Wieviele davon werden in Phrasenpaaren extrahiert?
- Für manche englischen Wortfolgen wird kein Phrasenpaar extrahiert. Suchen Sie dafür mindestens ein Beispiel.
- Wie beeinflussen nicht alignierte Wörter die Zahl der extrahierten Phrasen?
- Wieviele zusammenhängende Wortfolgen können aus einem Satz der Länge n extrahiert werden?

		0	1	2	3	4	5	6	7	8	9
		michael	geht	davon	aus	,	dass	er	im	haus	bleibt
0	michael										
1	assumes										
2	that										
3	he										
4	will										
5	stay										
6	in										
7	the										
8	house										

Implementieren Sie nun den Phrasenextraktionsalgorithmus auf der folgenden Seite. Der Algorithmus erhält drei Dateien mit Quellsätzen, Zielsätzen und symmetrisierten Alignierungen als Argumente und gibt die extrahierten Phrasen aus.

Erstellen Sie zum Testen drei Dateien mit dem obigen Alignmentbeispiel.

```

extract_all(e, f, A)
  BP := [ ] // extracted phrases
  for  $e_{start} \in [1, \dots, |e|]$  do
    for  $e_{end} \in [e_{start}, \dots, |e|]$  do
      // find the minimally matching foreign phrase
       $f_{start} := |f|$ ;  $f_{end} := 0$ 
      for all  $(e, f) \in A$  do
        if  $e_{start} \leq e \leq e_{end}$  then
           $f_{start} := \min(f, f_{start})$ 
           $f_{end} := \max(f, f_{end})$ 
      add extract( $f_{start}, f_{end}, e_{start}, e_{end}$ ) to BP

extract( $f_{start}, f_{end}, e_{start}, e_{end}$ )
  return [ ] if  $f_{end} = 0$ 
  // check if alignment points violate consistency
  for all  $(e, f) \in A$  do
    return [ ] if  $f_{start} \leq f \leq f_{end}$  and  $(e < e_{start} \text{ or } e > e_{end})$ 
  // add phrase pairs (including additional unaligned f)
  E := [ ]
   $f_s := f_{start}$ 
  repeat
     $f_e := f_{end}$ 
    repeat
      add phrase pair  $(e_{start} \dots e_{end}, f_s \dots f_e)$  to E
       $f_e := f_e + 1$ 
    until  $f_e > |e|$  or  $f_e$  aligned // i.e.  $\exists_{e'}(e', f_e) \in A$ 
     $f_s := f_s - 1$ 
  until  $f_s = 0$  or  $f_s$  aligned // i.e.  $\exists_{e'}(e', f_s) \in A$ 
  return E

```