Center for Information and Lang Processing
Prof. Dr. Alexander Fraser

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Erweiterungsmodul: Machine Translation **SS 2023**

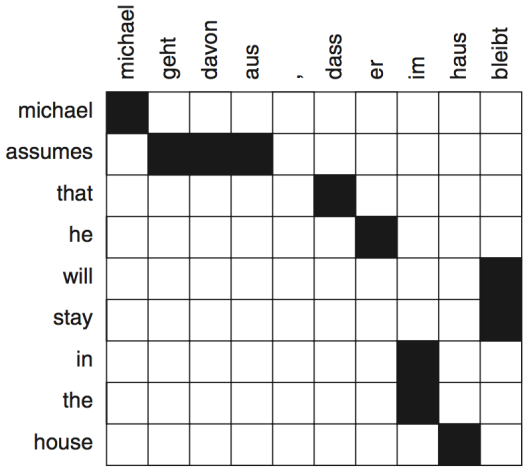
Exercise Sheet

Phrase Extraction, Language Models, Noisy Channel

(Thanks Costanza Conforti)
Please turn in your writeup as a PDF named as usual, like: `fraser_alexander_ex3.pdf`

Exercise 1. Phrase Extraction
Given the English-German sentence pair below, with the relative word alignment, extract all phrases consistent with the word alignment[1].

- How many distinct contiguous English and German phrases does this example contain? How many phrases are extracted?
- In some cases, given an English phrase, it is not possible to extract matching phrases in German. Find at least one example in this sentence pairs.
- Which is the effect of unaligned words in the number of extracted phrases?
- How many phrases can be extracted from a sentence of length n?




```
Input: word alignment A for sentence pair (e, f)
Output: set of phrase pairs BP
for e_start = 1 ... length(e) do
  for e_end = e_start ... length(e) do
    // find the minimally matching foreign phrase
    (f_start, f_end) = ( length(f), 0 )
    for all (e,f) ∈ A do
      if e_start ≤ e ≤ 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
function extract(f_start, f_end, e_start, e_end)
return {} if f_end == 0 // check if at least one alignment point
// check if alignment points violate consistency
for all (e,f) ∈ A do
  return {} if e < e_start or e > e_end
end for
// add pharse pairs (incl. additional unaligned f)
E = {}
f_s = f_start
repeat
  f_e = f_end
  repeat
    add phrase pair (e_start .. e_end, f_s .. f_e) to set E
    f_e ++
  until f_e aligned
  f_s --
until f_s aligned
return E
```

[1]Both the pseudo-code and the example in this exercise are taken from Koehn, Philipp. *Statistical machine translation*. Cambridge University Press, 2009

Exercise 2. Language Model, Noisy Channel[2]

(a) Calculate the parameters $p(e|e-1)$ of a Bigram Language Model from a corpus containing the following four sentences:

```
START the house is small
START the house is little
START the building is small
START the home building is small
```

(b) Given the sentence **f** = "das Haus ist klein" and the following parameters $p(f|e)$:

the		house		building		is		exists		little		small	
der	0.3	Haus	0.5	Gebäude	0.7	ist	0.7	ist	0.2	klein	0.7	klein	0.8
die	0.3	Heim	0.1	Haus	0.3	existiert	0.2	existiert	0.7	kurz	0.3	kurz	0.2
das	0.4	Gebäude	0.4			hat	0.1	hat	0.1				

calculate $p(e|f)$ of the following translations:

e ="the building is little"
e ="the house exists small"

using the Language Model of point (a) and IBM Model 1 as translation model. Recover that

$$argmax_e p(e|f) = argmax_e \frac{p(f|e)p(e)}{p(f)} = argmax_e p(f|e)p(e) \tag{1}$$

[2]Originally conceived by Laura Jehl, PhD candidate at the University of Heidelberg, modified by Costanza Conforti