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

LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Erweiterungsmodul: Machine Translation **SS 2023**

Exercise Sheet

Evaluating MT output

(Thanks to Costanza Conforti, Laura Jehl)
Please submit your writeup in a single PDF file called `lastname_firstname_ex4.pdf` (e.g., `fraser_alexander_ex4.pdf`).

Exercise 1. Given the following output translations of 2 MT Systems, with the corresponding reference[1]:

- System A - Israeli officials responsibility of airport safety
- System B - airport security Israeli officials are responsible
- Reference - Israeli officials are responsible for airport security

(a) Calculate the **Position-independent error rate** (PER) of both system.
Recall that

$$PER = 1 - \frac{correct - max(0, output\_length - reference\_length)}{reference\_length} \tag{1}$$

What do you observe?

(b) Then, calculate the **Word Error Rate** (WER) of both systems.
WER employs the *Levenshtein* distance, which is defined as the minimum number of editing steps needed to match two sequences. Considered editing steps are:

- *substitution* = replace one word with another
- *insertion* = add word
- *deletion* = drop word

To calculate the WER, proceed as follows: compute the Levenshtein distance by filling the matrix below, containing words from System B output on one axis and reference words on the other, as in has been done in the example below for System A output. Remember that the Levenshtein distance between two strings of length $|a|$ and $|b|$ is given by:

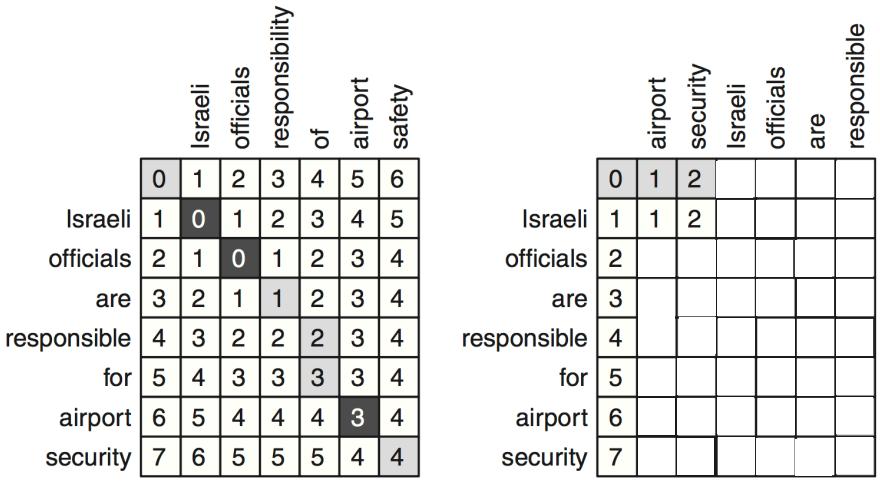
Given the Levenshtein distance, calculate the WER of both systems using the formula:

$$WER = \frac{substitutions + insertions + deletions}{reference\_length} \tag{2}$$

What do you observe?

[1]The examples of this exercise, as well as the Figure, are taken from Koehn, Philipp. *Statistical machine translation*. Cambridge University Press, 2009

$$\operatorname{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \operatorname{lev}_{a,b}(i-1,j) + 1 \\ \operatorname{lev}_{a,b}(i,j-1) + 1 \\ \operatorname{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$



(c) Finally, calculate the **BLEU** score of both sentences.

BLEU score combines n-gram precision with a *brevity penalty*, defined as $min(1, \frac{output\_length}{reference\_length})$. Using the formula:

$$\text{BLEU-N} = min(1, \frac{output\_length}{reference\_length}) \prod_{i=1}^{n} precision_i \tag{3}$$

calculate the BLEU score with maximum order n for n-grams to be matched equal to 1, 2, 3 and 4. (Note that BLEU-4 is standardly used in the literature)

Metric	System A	System B
precision (1-gram)	/6	/6
precision (2-gram)	/5	/5
precision (3-gram)	/4	/4
precision (4-gram)	/3	/3
brevity penalty (4-gram)	/7	/7
BLEU-1		
BLEU-2		
BLEU-3		
BLEU-4		