NAME: XIANG LI, GWid: G47754486

Bonus Homework #4

Instructions:

- ✓✓ This assignment is due on **Tuesday November 19, 2019, by 11:59 pm**.

a) [1 point] Apply forward alignment and draw the word alignment matrix for the first pair of sentences:

```
English: let us all strive to live and let live .
French : employons-nous tous à vivre et à laisser vivre .
```

A: 0-0 2-1 4-2 5-3 6-4 4-5 7-6 8-7 9-8

b) [1 point] Apply the reverse alignment and draw the alignment matrix for the same pair of sentences above.

A: 0-0 1-0 2-1 3-0 4-3 5-3 6-4 7-6 8-7 9-8

c) [1 point] Apply the summarization tool (atools) to symmetrize the forward and reverse alignments. Draw the final alignment matrix.

A: 0-0 1-0 2-1 3-0 4-2 5-3 6-4 7-6 8-7 9-8

d) [2 points] Using a large parallel corpus and a word alignment tool like `fast_align` (plus any additional post-processing), describe how you could extract a set of **paraphrases** for English. Paraphrases are different phrases in the same language that have similar meaning, for example:

- "symptoms of influenza include fever"
- "elevated temperature is a sign you have the flu"

A: Considering the functionality of tool fast_align, "it can produce output in the widely-used i-j "Pharaoh format" where a pair i-j indicates that the *i*th word (zero-indexed) of the left language (by convention, the *source* language) is aligned to the *j*th word of the right sentence (by convention, the *target* language)."

Then, a very simple and straightforward idea is to pick up sentences from another language and using a large parallel corpus in our processing English text for word alignment. And, collecting all the good alignment of these parallel corpus because they are highly likely to be the **paraphrases**. And we can use minimum distance to quantify the goodness of alignment between your processing English text and one sentence from other language.

Reference:

1. Chris Dyer, Victor Chahuneau, and Noah A. Smith. (2013). A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proc. of NAACL*.