CSCI 3907/6907: Introduction to Statistical NLP

Bonus Homework

Instructions:

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

Word Alignment

Download and install `fast_align` (https://github.com/clab/fast_align), which is an efficient implementation of IBM Model 2 (a small extension of IBM Model 1 covered in class). Follow the instructions in the page above to answer the following questions.

Remember that in word alignment models, each target word is aligned to exactly a single source word. Symmetrization is used to get many-to-many word mappings. Using the provided parallel dataset (`en_fr.txt`):

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 .
```

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

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

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"