Machine Translation Problem Set

Due: Tuesday Mar. 1st, 11:00am (Eastern Time)

Typeset your solution in a file called `lab3-pset.tex`, and push both the source LaTeX file and the corresponding `lab3-pset.pdf` to the Lab3 repository under the `pset` directory. The goal here is to get you thinking about the properties of our MT algorithm before you actually start implementing anything.

Q1) Linguistic Fieldwork

Consider the following English sentences:

The dog drank the coffee.
The student drank the soup.
The coffee warmed the student.

We have a parallel corpus in Fthishr, a language that loves its coronal fricatives and word-final alveolar approximants.

Thir o favezh or shath.
Sirzh o zhiidh or shath.
Favezh or sirzh o vozir.

Give the correct alignment for the following pair of sentences:

The soup warmed the dog.
Zhiidh or thir o vozir.

Q2) Expectation-Maximization

Compute the values of $n_{\text{'eat','mange'}}$ and $\tau_{\text{'eat','mange'}}$ after the first and second iterations of EM given the following training data:

Elle mange du pain	*She eats bread*
Il mange du boef	*He eats beef*

Q3) Reasoning about partial-counts

In class (and in the reading: section 2.2.1), we noted that $n_{e,\circ}$ may not be the same as the number of times the word e appears in the training corpus. Explain.

Q4) Reasoning about alignments

Would $< 0, 0, 0 >$ be a legal alignment for a French sentence of length three? Describe some implication of this fact.

Q5) Reasoning about EM

On the first iteration of IBM model 1 training, the word that would align the most often with *'pain'* is almost certainly going to be *'the.'* Why?

Yet in the Figure 2.6 from the reading (copied here below) it does not mention *'the'* as a translation of *'pain.'* Why is that?

English word	Iteration 1	Iteration 2	Iteration 19	Iteration 20
bread	0.042	0.138	0.3712	0.3710
drudgery	0.048	0.055	0.0	0.0
enslaved	0.048	0.055	0.0	0.0
loaf	0.038	0.100	0.17561	0.17571
spirit	0.001	0.0	0.0	0.0
mouths	0.017	0.055	0.13292	0.13298

Figure 2.6: Probabilities for English words translating as '*pain*'

Bonus (Extra Credit)

I suggested starting EM with all $\tau_{e,f}$'s the same. In fact, as long as they are the same (and non-zero) any one value works as well as any other. That is, after the first M-step the τs you get will not be a function of the τs you stated with. For example, you could initialize all τs to one. Prove this.

However, although the τs at the end of the M-step would not vary, there would be some other variable inside the EM algorithm that would. What is it and why?