CMPE 409 - Machine Translation

Middterm Project

Deadline: 23:00, May 17

In your assignment,explain your codes with *comments*. Without *comments*, your assignment will not be marked.

Problem

In this assignment you are asked to implement the IBM Module I (Figure I) with python programming language.

Input: set of sentence pairs (**e**, **f**)
Output: translation prob. $t(e|f)$

initialize $t(e|f)$ uniformly
while not converged **do**
  // *initialize*
  count$(e|f) = 0$ **for all** e, f
  total$(f) = 0$ **for all** f
  for all sentence pairs (**e**,**f**) **do**
    // *compute normalization*
    for all words e in **e** **do**
      s-total$(e) = 0$
      for all words f in **f** **do**
        s-total(e) += $t(e|f)$
      end for
    end for
    // *collect counts*
    for all words e in **e** **do**
      for all words f in **f** **do**
        count$(e|f)$ += $\frac{t(e|f)}{\text{s-total}(e)}$
        total(f) += $\frac{t(e|f)}{\text{s-total}(e)}$
      end for
    end for
  end for
  // *estimate probabilities*
  for all foreign words f **do**
    for all English words e **do**
      $t(e|f) = \frac{\text{count}(e|f)}{\text{total}(f)}$
    end for
  end for
end while

Figure 1: EM training algorithm for IBM Model 1

In your code test the following tasks:

- In each iterations print out:

  - Iterate total five times over the two pairs sentence that we have explained in the lecture
  - s-total(e) values for each pairs
  - expected counts: count(e|f)
  - total(f)
  - estimate probabilities: t(e|f)

- Test your code with five pairs Turkish-to- English sentences (5 parallel sentence, you can write these sentence yourself). Just record result of the last iteration
- Compare your results with the Python IBM modules of the NLTK library
- Write a short report that contains the result that you have gotten from previous tasks.

Submission

- Submit your source code with **a readme.txt** file. Your code should include comments. Be sure you have understand it.
- Submit your report: It contains the data you have gotten from in previous section.

Hint: Look at the lecture notes and examples as references