NLA Assignment 1 - Word Alignment

Harshita Sharma - 20171099

IBM Model 1

Data preprocessing:

- The datasets are already tokenised.
- All words are changed to lower case.
- 'NULL' is added to source and target sentences.

Training the model:

- Approach: The following pseudocode has been used in order to implement IBM Model 1 in Python:

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

- Used defaultdict as a translation probability table to improve training time and space, where each entry takes key-value pairs in the following format: tef([hindi_word, english_word]) = translation_probability.
- Using this only relevant pairs of words are looked at.

- Each Hindi word in each Hindi sentence is paired with the corresponding English translated sentence's words instead of using all possible word pairs from the entire dataset.
- The EM algorithm is run i.e. the model is trained for 15 epochs.

Output Analysis:

- Top 20 pairs with the highest probabilities are:

  P(रिंक|skatin) 1.0
  P(महादेव|mahadeva) 1.0
  P(गंगा|ganges) 0.99
  P(राजस्थान|rajasthan) 0.99
  P(भूटान|bhutan) 0.99
  P(भवन|bhavan) 0.99
  P(दिल्ली|delhi) 0.99
  P(ऐतिहासिक|historical) 0.99
  P(महादेव|mahadev) 0.99
  P(हजार|thousand) 0.99
  P(आज|today) 0.99
  P(बस्तर|bastar) 0.99
  P(पटना|patna) 0.99
  P(मिजोरम|mizoram) 0.99
  P(धर्मशाला|dharamshala) 0.99
  P(नेहरू|nehru) 0.99
  P(नदी|river) 0.99
  P(जैन|jain) 0.98
  P(नाम|name) 0.98
  P(बंगाल|bengal) 0.98

Error Analysis:

- Since IBM model 1 does not take word order into consideration - as no LM is used - the errors generated are substantial.
- IBM Model 1 is weak in terms of conducting reordering or adding and dropping words.