LIN570: HW11 – mt (up to 100pts)

YOUR NAME (`UW NetID`)

Due date: 11pm on Dec 12, 2019 (Tuesday)

All the example files are under `~/dropbox/19-20/570/hw11/examples/`.

1. **Q1 (40 points)**

   (a) (10 points): prepare your data including removing xml tags, senetence boundary detection, tokenization. See `tools` at `https://www.statmt.org/europarl/v7/tools.tgz` for preprocessing.

   - your preprocessed files will be: `en-ep-99-12-17.tok.txt` and `de-ep-99-12-17.tok.txt`

   (b) (30 points): implement a sentence aligner using the Gale and Church algorithm:

   - `./setence_aligner.sh de-ep-99-12-17.tok.txt en-ep-99-12-17.tok.txt > de-en-aligned.txt`
   - `cut -f1 de-en-aligned.txt > de-en-aligned.txt.de`
   - `cut -f2 de-en-aligned.txt > de-en-aligned.txt.en`

2. **Q2 (40 points):**

   (a) (10 points) discuss how to evaluate sentence aligned results intrinsically (recall evaluation on sentence boundary detection).

   (b) (30 points) implement `eval_sentence_alignment.sh`.

   - `./eval_sentence_alignment.sh ep-99-12-17-de-en.de ep-99-12-17-de-en.en de-en-aligned.txt.de de-en-aligned.txt.en`

3. **Q2 (20 points):** show the MLE probability parameters (M-step) by normalizing the counts to sum to 1 (i.e., $t(f|e) = \frac{count(f|e)}{total(e)}$) after the second iteration: (See MT slides)

$$
\begin{array}{rrr}
t(maison|green) = & t(vert|green) = & t(la|green) = \\
t(maison|house) = & t(vert|house) = & t(la|house) = \\
t(maison|the) = & t(vert|the) = & t(la|the) =
\end{array}
$$

The submission should include:

- The `readme.`[`txt`|`pdf`] file includes answers for Q2a and Q3.
- `hw.tar.gz` includes
  - `setence_aligner.sh`
  - `de-en-aligned.txt.en`
  - `de-en-aligned.txt.de`
  - `eval_sentence_alignment.sh`