CS378 Assignment 5: Alignment and Machine Translation

Due date: Thursday, November 12 at 11:59pm CST

Academic Honesty: Please see the course syllabus[1] for information about collaboration in this course. While you may discuss the assignment with other students, **all code you write and your writeup must be your own!**

Goals There are two goals for this assignment. First, you will study what IBM Model 1 does in a Spanish-English alignment setting to better understand what correspondences it learns between these two languages. Second, you will look at attentions in a Spanish-English neural machine translation system to understand how the properties of those "alignments" are similar to and different from Model 1.

Dataset and Code

Please use Python 3.5+ for this project.

Data The dataset for this assignment is the Europarl dataset.[2] This is a large dataset of professionally translated sentences from the proceedings of the European Parliament. You will be focusing on English to Spanish word alignment (that is, aligning each target English word to one or more source Spanish words that are associated with it) and Spanish to English translation.

Part 1: Word Alignment (30 points)

In this part, you will experiment with the IBM Model 1 alignment model as discussed in lecture. Recall that these models take a bitext sentence as input: a pair of sentences $\mathbf{w}^{\text{en}}$ (length n) and $\mathbf{w}^{\text{es}}$ (length m) that are translations of each other. For aligning English to Spanish, we augment $\mathbf{w}^{\text{es}}$ to additionally have a special `NULL` token at the end, making this sequence of length $m + 1$.

The alignment models discussed in lecture (Model 1 and the HMM) have the form:

$$P(\mathbf{w}^{\text{en}}, \mathbf{a}|\mathbf{w}^{\text{es}}) = P(\mathbf{a}) \prod_{i=1}^{n} P(w_i^{\text{en}}|w_{a_i}^{\text{es}})$$

where $w_{a_i}^{\text{es}}$ is the a_ith token in the Spanish sentence. a_i controls what Spanish word the English word "decides" to condition on to be produced. When aligning English to Spanish, we learn a matrix of word translation probabilities $P(w^{\text{en}}|w^{\text{es}})$, the probability of producing an English word given a particular Spanish word. In this project, we use IBM Model 1, which sets $P(\mathbf{a}) = \prod_{i=1}^{n} \frac{1}{m+1}$, a constant uniform distribution, meaning the only model parameters are the translation probabilities.

Getting started The main command to run is:

```
python model1_aligner.py
```

[1] http://www.cs.utexas.edu/~gdurrett/courses/fa2020/syllabus.shtml

[2] http://statmt.org/europarl/

This loads the first 10,000 lines of the data, tokenizes and indexes the data, filters to keep only Spanish sentences of length at most 15, and trains IBM Model 1 on this data using 10 iterations of the expectation maximization (EM) algorithm. This algorithm optimizes $\sum_{i=1}^{D} \log \sum_{\mathbf{a}} P(\mathbf{w}^{\mathrm{en(i)}}, \mathbf{a}|\mathbf{w}^{\mathrm{es(i)}})$, the marginal log likelihood of generating the target given the source.[3] With a trained model, we can do posterior inference: `infer_model_1` computes the posterior distributions $P(a_i|\mathbf{w}^{\mathrm{en}}, \mathbf{w}^{\mathrm{es}})$ for each i. In Model 1, the a_i can be handled independently.

Because we do not have labeled alignment data, we cannot directly evaluate the model's alignment performance. However, we can still manually look at what the model is doing. The code prints alignment posterior probabilities on the training data; note that because the data is unlabeled, we have no need of a separate train/test dataset split. By default, the code will pretty-print output to the terminal. If you add the `--make_vis` flag, the model will produce alignment figures as PDFs and save these for the first 10 examples in the dataset; `--show_plot` will additionally display these figures.

Q1 (8 points) Identify an instance where a single token in English is aligned to two non-NULL tokens in the corresponding Spanish sentence (probabilities greater than 0.3) and this alignment is **incorrect** by your judgment. (a) List the sentence pair and the two Spanish words with high probability, and (b) describe what Model 1 did here and why it could be happening. (Use Google Translate if you don't know what words mean.) **Note that an alignment could still be okay even if the translation is, strictly speaking, incorrect – use your judgment about what is going on!**

Q2 (8 points) Identify an instance where a single token in English is aligned to two tokens in Spanish and this alignment is **plausible** (seems linguistically correct). Describe what is happening here and why.

Q3 (7 points) There is a phenomenon called *garbage collection* in alignment. In these cases, many words on the target side are aligned to the same source-side word. Find an occurrence of garbage collection in the data: at least three English words aligned to the same Spanish word with probability greater than 0.5, and some of them are erroneous. **Describe the two sentences involved and what you observe.** (Note that this is asking about a different case than Q1, which was asking about one English word aligning to many Spanish words.)

Q4 (7 points) Suppose we have vocabularies consisting of Spanish words s_1 and s_2 and English words e_1 and e_2 and are aligning English to Spanish as we've been doing. Suppose that we have the following translation matrix $P(e_i|s_i)$:

	e_1	e_2
s_1	0.9	0.1
s_2	θ	$1-\theta$
NULL	0.5	0.5

a) Suppose our corpus consists of a single sentence pair with source s_2 and target e_2. What value(s) of θ maximize the marginal likelihood $\sum_{\mathbf{a}} P(\mathbf{w}^{\mathrm{en}}, \mathbf{a}|\mathbf{w}^{\mathrm{es}})$ of this sentence under Model 1? If multiple values or a set of values all lead to likelihood being maximized, describe the set of values.

Note that you don't need to know the EM algorithm to do this problem. Hint: write down the marginal likelihood by explicitly summing over alignments (there aren't that many possibilities on a short example).

[3] The EM algorithm is used because this is a *nonconvex* optimization problem due to the sum inside the logarithm. The problem is therefore significantly harder than the supervised parameter estimation we did for HMMs, where we just counted and normalized the data.

b) Suppose our corpus consists of a single sentence pair with source $s_1\ s_2$ and target $e_1\ e_2$. What value(s) of θ maximize the marginal likelihood of this sentence under Model 1? If multiple values or a set of values all lead to likelihood being maximized, describe the set of values.

Part 2: Attention in Machine Translation (20 points)

Second, you will conduct a similar analysis on some machine translation output. The data is again taken from Europarl. We use 1,572,587 sentences for training (not given to you) and tested on a 1000-sentence development set (`europarl-dev.en` and `europarl-dev.es`) that doesn't overlap with the training as the development set. The model we used is an LSTM with attention from OpenNMT. Training this system is relatively time-intensive, taking a few hours on a 1080 Ti GPU.

We have given you an output file `europarl-esen-attns.txt` with 1000 decoded sentences and their visualizations. Each sentence and its prediction is listed on their own lines. The ground-truth English sentences are included in a different file. The attentions are shown as a matrix with the attention distribution for each English word (which corresponds to a single decoder timestep) shown on a separate line. The highest attention value is marked with a star. Note that the English and Spanish words are both abbreviated for this visualization, so longer words will be truncated.

Q5 (5 points) Look at the second sentence in the development set. Identify **two words** in English that seem to be aligned to the "wrong" word in Spanish on the basis of attention. (a) What words are these? (b) Why do you think the model is able to produce the correct word despite the seemingly incorrect alignment?

Q6 (7 points) Identify a translation which features out-of-order alignments; that is, the alignment pointer "moves back" at some point during decoding. Give the translation, and say whether you think this out-of-order alignment reflects a real linguistic divergence or is just noise, an error, etc.

Q7 (7 points) Identify **two translations** that you believe contain translation errors. Give the two translations and discuss (a) what the errors are; and (b) how bad you think these mistakes are in terms of harming comprehension of the sentence.

Deliverables and Submission

You only need to submit a writeup for this assignment. Please submit on Gradescope.