Assignment #5
Due April 25th

Machine translation

About

In this assignment you'll gain experience training a simple sequence to sequence model. Originally, we were going to work with structured data (which is a far more common and essential skill), but I think beyond what we cover in class - there's not much to add. So, we'll do this instead - it's amazing that it's possible with such a small amount of code.

Submission instructions

Please submit this assignment on CourseWorks. You submission should include one or more Jupyter notebooks.

References

- [8-seq2seq.ipynb](#) starter code.
- Here's a handy [dataset](#) for English to Spanish.

Part 1 (40 points)

Train a model to translate between two languages of your choice (say, from English to Spanish) using a couple thousand sentences.

Part 2 (20 points)

Train a second model to translate between the same two languages in reverse order (say, from Spanish to English).

Part 3 (40 points)

Back-translate. Use your two models to translate a sentence from English to Spanish, and then back to English. Compare the original sentence, and the back-translated sentence. Repeat this using an evaluation corpus of 1,000 sentences, and report the BLEU score.

Extra credit

You may complete some of all of these problems, in any order.

EC1: Investigate interlingual representations.

Train an encoder-decoder model to translate between two source languages (say, English and Spanish) into one target language (say, French). Once the encoder is trained, use it to encode similar sentences in the source languages (e.g., "Rachel likes coffee" and "Rachel le gusta el cafe"). Investigate how the encoder compresses and represents concepts from the source sentences.

Can you find any elements in the encoded vector that correspond to specific concepts? If you use a small corpus with a tiny number of concepts, and a tiny encoding vector, it may be possible to find something cool.

EC2: Attention mechanisms.

Implement and compare two different attention mechanisms. How do they affect your BLEU score in part three?