Quick Task

Please submit by Saturday, Feb 6. Thanks!

Please submit a one page PDF describing your approach and results. Be concise in your report and emphasize clarity. It is more important that your report will be clear and work well reasoned than picking a fancy technical approach. Please include a link to a Github repository containing your code, and make sure I can access it. You can email your PDF directly to yoav@cs.cornell.edu.

This is a short classification task, focused on machine translation. Instead of actually building a system that does machine translation, you will build a classifier that can tell whether a translation was created by a human or by a machine.

The data are divided into two files: a training set and a test set. Each is formatted so that it includes a source sentence in Chinese, a human translation of that sentence to English (called the *reference*), another translation to English (either by a machine or human, called the *candidate*), a score for the quality of the translation,[1] and a label indicating whether the candidate comes from a machine (`M`) or human (`H`).

The data are encoded in UTF-8, since they include Chinese characters. You will probably want to use Python's support for dealing with UTF-8 text in this assignment.

Your task is to build a classifier that tells whether a candidate is a human or machine translation. In addition to the candidate, your classifier may consider the source sentence and the reference. The Bleu scores provided may also be used as inputs to your classifier, as well as more complex versions of Bleu or other MT evaluation scores you choose to calculate. You are welcome to use any existing resources, tools, or libraries to build your classifier. You may even use additional data, with the exception of the test data.

Please evaluate using the average of F_1 for the human and machine classes.[2] Please implement your own evaluation script, or use one from a third-party tool.

[1] he score is a very simple version of Bleu that considers only unigrams. Bleu which is a common machine translation metric. If you are interested, you can read about it here. The script for computing Bleu scores can be found on the NIST website. You will need to reformat the data if you want to run this script. You can easily find Python implementations as well, although they will likely include a more complete version of the score.

[2] You should be able to see why a most-frequent-class classifier won't do well on this score.