Problem Set 1

- Hang Gao (uni: 2469)
- Email: hang.gao@columbia.edu
- Date: Sep 24, 2017

Question 4

Setup

```
> cd ./pset1
> chmod +x ./run
> ./run q4

**** With total running time of 1.59s
Done. Please check q4.log for evaluation.
```

Result

- `ner_train.merged.dat`: the merged training corpus by clipping infrequent words.
- `ner_dev.baseline.pred`: the prediction file with log probability.
- `q4.log`: the performance log file.

Evaluation

```
> cat q4.log

Found 14043 NEs. Expected 5931 NEs; Correct: 3117.

         precision      recall          F1-Score
Total:   0.221961       0.525544        0.312106
PER:     0.435451       0.231230        0.302061
ORG:     0.475936       0.399103        0.434146
LOC:     0.147750       0.870229        0.252612
MISC:    0.491689       0.610206        0.544574
```

Observation

- At this section, we trained a baseline model for the NER problem. Not Surprisingly, the performance is relative weak - at an average F1-Score of

`0.31.`

- From a high level, the recall is much higher than precision, indicating that the model does relatively well in the most retrieving positive tags, but producing a lot of false alarm.
- Note that `LOC` category has the lowest precision and highest recall, which in turn produces the lowest F1-Score. It means that the false alarms are particularly severe in given tagging task, which is perfectly reasonable since the location and person's name can be hard to detach without the grammer information.

Question 5

Setup

```
...

> ./run q5

**** With total running time of 17.65s
Done. Please check q5.log for evaluation.
```

Result

- `ner_train.merged.dat`: the merged training corpus by clipping infrequent words.
- `ner_dev.trivit.pred`: the prediction file with log probability.
- `q5.log`: the performance log file.

Evaluation

```
> cat q5.log

Found 4704 NEs. Expected 5931 NEs; Correct: 3640.

         precision      recall          F1-Score
Total:   0.773810       0.613724        0.684532
PER:     0.757660       0.591948        0.664630
ORG:     0.611855       0.478326        0.536913
LOC:     0.876458       0.696292        0.776056
MISC:    0.830065       0.689468        0.753262
```

Observation

- First note that the most improvement is that we now have much less alerts of NEs (14043/4704) which means our precision is much higher than before. This improvement, in my opinion, is close related to the trigram MLE and HMM's contextual information. Since in the baseline model, we cannot access to the contextual knowledge, so that we have to draw same prediction for a given word. But now, since we are not constrained by local information, a better understanding of tagging could be formed.