Evaluating tagger output

This page provides more information to help you compare the tagger output. We have provided a perl program (compare_taggers.pl) that can be used to compute the confusion matrices and kappa for the taggers, comparing them against the 'truth' in the .pos files. For a definition of Kappa, please see the class notes or the Jurafsky text; Kappa ranges between 0 (for no agreement between the tagger and the Gold Standard) and 1 (for complete agreement); typically excellent values are greater than 0.8. The program will also print the P(A) and P(E) values used to compute kappa (recall that P(A) is the agreement between 'truth' and the tagger, while P(E) is the agreement expected by chance alone.)

Running the program

The basic syntax for compare-taggers.pl is as follows.

```
prompt%> compare-taggers.pl (-b|-h) (-m) (-k) tagger-output gold-
standard
```

The parameters are as follows:

tagger-output
: The file containing the output from the tagger

gold-standard
: The file containing the "gold standard" tags

-b
: Specify that the output is from the Brill tagger

-h
: Specify that the output is from the HMM tagger

-m
: Print out the confusion matrix

-k
: Print out kappa

Examples

To print the confusion matrix and kappa for the tagged file wsj_1975.brill, tagged by the Brill tagger, and save the matrix and kappa in the file wsj_1975.results:

```
prompt%> compare-taggers.pl -b -m -k wsj_1975.brill wsj_1975.pos >
~/wsj_1975.results
```

To print kappa only for the tagged file sw2019.hmm, tagged by the HMM tagger:

```
prompt%> compare-taggers.pl -h -k sw2019.hmm sw2019.pos
```

Notes

- Please tag the "raw" files (with the nasty formatting) when you are going to use compare-taggers.pl - it relies on the formatting being the same in the tagged and gold-standard files. You will know you have gone wrong if you get a lot of "line mismatch" errors, and a low kappa score (near 0).
- To compute the confusion matrix or kappa for all of the texts, use the files all.raw and all.pos, which contain all of the texts in one big file.
- The confusion matrix is an n x n matrix of tag names across the columns and down the rows, with a count of the number of times that one tag was (incorrectly) substituted for another. If nothing appears in a cell, the count is zero. It is most often used in speech recognition testing, but it is just as useful here to pick out which systematic incorrect substitutions are being made.
- There is nothing special about this perl program, so you should be able to copy it to your own directory and use it on Linux boxes too.