Week 3 lab

Beginning annotation for assignment 1

Having designed an annotation schema, we manually label a corpus with three goals in mind:

1. assessing and refining the categorisation schema
2. evaluating an automatic classifier
3. training a supervised machine learning classifier

Provided with the scheme for Assignment 1, everyone will begin independently labelling the same set of documents. After 10 minutes, divide yourselves into pairs and discuss any disagreements with your partner. The class will discuss the kinds of disagreements raised.

You should continue your annotation independently for the assignment, noting that the common portion should be completed by Monday.

Evaluation and error analysis for binary classification

Two assignments of categorical labels for the same items (e.g. two human annotators or a gold standard and a system prediction) can be compared using a number of evaluation metrics, including Precision, Recall, F-score and Cohen's kappa. All of these metrics can be derived from a confusion matrix which shows, for each pair of categories `c` and `d`, how many items were placed in category `c` by annotator `A` and category `d` by annotator `B`.

Using a spreadsheet (or your tool of choice), we will experiment with how the overall performance metrics change with respect to the underlying confusion matrix. We can start with a binary confusion matrix, i.e. a 2x2 grid. Let's assume we have 100 instances, with 80 in the Negative category and 20 in the Positive category.

We first construct the confusion matrix where both annotations are identical, and declare the row labels those of the unchanging gold standard.

Go to **Google Sheets** (https://docs.google.com/spreadsheets/d/1fbdRJMqd9oZKNB8wko6SZSCksn4uYa161nw-SRyAwPY/edit) and choose File: Make a Copy.

Gold↓ \ Predicted→	Negative	Positive	Total gold
Negative	80	0	80
Positive	0	20	20
Total predicted	80	20	100

Exercises

Please work in pairs or groups of three

1. In other cells on the spreadsheet, calculate:
   - Accuracy (the proportion that are agreed)
   - Precision, recall and **F-score** **(http://en.wikipedia.org/wiki/F1_score)** for each class
   - **Cohen's kappa** **(http://en.wikipedia.org/wiki/Cohen%27s_kappa)**

2. (Feel free to implement other metrics too. "Balanced accuracy", the average of recall in both/all classes, has recently become popular in some fields for binary classification tasks. You may also implement Fleiss' kappa; while it is rarely used for two annotators, you will notice that it differs from Cohen's kappa.)

3. Now create some confusions/disagreements in the confusion matrix and watch how the metrics are affected. Make sure that the total number of items is unchanged, and the 80:20 distribution of classes in one annotator (the "gold standard") is unchanged.

   Explore how the metrics respond to different kinds of errors ("true negatives" and "false positives"):

   1. where precision and recall are similarly moderate
   2. where precision and recall are dissimilar
   3. where precision or recall is 0
   4. where kappa is negative, 0 and maximised
   5. with errors distrubuted as if the predictions were random. (If, for instance, we randomly assigned 50% of all samples the positive class, what is the number of true positives, true negatives, false positives, false negatives? What if we randomly assigned 10% the positive class?)

4. Sometimes we care about all classes equally. Sometimes there is a majority class which we don't care about (e.g. when trying to determine if a blood test is able to identify some pathology, we care less about the cases that are negative and predicted as negative than we do about the other cells of the confusion matrix). How does this affect your choice of metric?

5. When analysing agreement and disagreement, you often want to get some understanding of why you disagreed. How do you go from metrics to looking at informative examples? Which metrics help you identify examples worth investigating?

6. If you were reporting these metrics, how many decimal places (or significant figures) would be appropriate to report?

7. When evaluating an automatic classifier, there is usually a ground truth against which a prediction is compared. In human annotator agreement, there is usually no ground truth, just a series of

8. Advanced: How might the choice of metric change when it is an annotator agrement measure vs. when one is the ground truth and the other is a system's classification?
9. Advanced: In information retrieval and information extraction tasks, we presume the number of true negatives (the number of documents you have correctly *not* retrieved, the number of phrases you have correctly *not* recognised as a person's name) is very large. How does this affect the choice of metric?

Feel free to now modify the number of instances in each class, e.g. a more balanced distribution (50:50), more imbalanced (95:5) or imbalanced in the other direction (20:80). How does the *prevalence* of the True class affect the measures? How useful is each metric for evaluating a class that has very few instnaces?

While it is valuable to understand the behaviours of each metric, and the trade-offs they represent, a good error analysis always requires more investigation than just reporting measures. For an analysis, the distribution of errors would be fixed; you would use the metrics and confusion matrix to identify interesting or problematic parts of your data, and qualitatively investigate and reason about (attempt to explain) the kinds of agreements and disagreements being produced.