Homework #2: Text summarization

Assigned: October 25, 2002
Due: November 8, 2002

This assignment covers text summarization. You will be using an existing summarizer, mead, available on the clair machines. To get started, you have to add the path to mead to your PATH: /clair4/projects/mead307/stable/mead/bin/.
The documentation for mead is available from the Web at http://tangra.si.umich.edu/clair-/mead. You only need to download mead itself from the web page if you plan to run it on a different machine (not recommended for this assignment).

1. (10 points) Copy the files from /clair4/class/lni-f02/data-hw2 to a directory where you have write access. Let’s call that directory data-hw2. Inside that directory, you should now have a subdirectory milan which contains a file, milan.cluster. That file indicates that the milan cluster consists of nine documents, all of which are located in a specific format in the docsent subdirectory.

2. (10 points) Go to your data-hw2 directory. Run mead as follows:
   mead.pl -extract -s -p 10 milan > milan.10.extract
   This will produce a sentence extract consisting of 10% of the sentences in the original cluster.
   Run mead once more, this time specifying the -summary option.
   mead.pl -summary -s -p 10 milan > milan.10.summary
   This will produce a summary corresponding to the sentences in the original cluster.
   DO NOT read this summary yet. Just make sure that it is produced properly.
   Submit both the automatic extract and the automatic summary.

3. (10 points) Manually create a ten percent extract of the milan cluster. Use the same format for your extract as the automatic extract above. Your extract should indicate which sentences from the original cluster you consider the most important to tell someone what happened. In selecting these sentences, pay special attention to facts that change with time. You probably want to extract sentences that indicate the latest version of the facts that is available in the nine articles.
   Run extract-to-summary.pl to produce the summary equivalent to your manextract.
   Submit your manextract file and your manual summary file.
4. (10 points) Run meadeval.pl to compare your extract with the manually produced extract.

   meadeval.pl milan.10.extract milan.10.manextract

   Submit the output of meadeval.pl.

5. (10 points) Run mead.pl in RANDOM and LEAD-BASED mode and save the resulting extracts as milan.10.randextract and milan.10.leadextract.

   Submit these two files.

6. (10 points) Run meadeval.pl to compare the random and lead extracts with your manual extract.

   Submit the output of meadeval.pl for each of these two cases.

   Discuss these numbers in contrast with the numbers comparing mead’s output and the manual extract. Are any of these numbers higher or lower than what you’d expect? Why?

7. (10 points) The automatic summary was produced using the mead default configuration. Read the section on .meadrc in the mead documentation. Modify your .meadrc file by varying the parameters passed to the classifier and reranker. Pay special attention to the specific semantics of the Length parameter.

   Submit three meadrc files that perform better than the default version of mead. For each of them, submit the resulting extracts, summaries, as well as the output of meadeval.pl. Note: if you cannot outperform mead, you should still submit three meadrc files and the corresponding results.

8. (10 points). Explain how you selected the values in the previous item. Why (or why not) were you able to outperform the default configuration of mead. How could one build a system that can learn the right values of the meadrc parameters to handle different types of input (e.g., news vs. other types of text, different input size, etc.)

9. (10 points) Pick a cluster from www.newsinessence.com (NIE) - the cluster has to contain at least three or four articles in it from at least two sources. If a summary hasn’t yet been created, get NIE to generate it first.

   Submit a printout of the NIE-generated cluster and summary and indicate the cluster ID from which the summary was created.

10. (10 points) Find five problems in the summary (not the extract) produced by mead. Think how one could design a version of mead that avoids such problems. For each problem, indicate a possible reason and a possible solution.