Opinosis Opinion Dataset 1.0 - Documentation

Last revised: Tuesday, July 20, 2010
Author: Kavita Ganesan (kganes2@illinois.edu)

Dataset Overview

This dataset contains sentences extracted from reviews on a given topic. Example topics are *"performance of Toyota Camry"* and *"sound quality of ipod nano"*, etc. There are 51 such topics in this dataset and the opinions are obtained from Tripadvisor(hotels), Edmunds.com(cars) and Amazon.com(various electronics). There are approximately 100 sentences per topic. This dataset was used for the following paper:

*Kavita Ganesan, ChengXiang Zhai, and Jiawei Han, "**Opinosis: A Graph Based Approach to Abstractive Summarization of Highly Redundant Opinions**", Proceedings of the 23rd International Conference on Computational Linguistics (COLING 2010), Beijing, China, 2010.*

The Opinosis dataset also comes with human composed summaries used for the above paper. I have also provided some scripts to help with the summarization/evaluation tasks using ROUGE. For any questions, contact kganes2@illinois.edu.

Key idea of Opinosis

This paper presents a flexible framework for generating very short abstractive summaries. The key idea is to use a graph data structure referred to as the Opinosis-Graph to represent the text to be summarized and then find paths through this graph to produce concise abstractive summaries. While the evaluation is on an opinion dataset, the approach itself is general and can be applied to any dataset that contains high amounts of redundancies for example twitter data or user comments.

The Dataset

Assuming that you have unzipped the dataset file, you will see the following folders**:**

Folder	Content Description
topics/	• This is where the topic based sentences reside. • Each file with the '.data' extension corresponds to a topic and the filename actually describes the topic. Each of these files will contain a set of sentences relevant to the topic. • Note: Due to imperfect sentence segmentation (from the original text), there may be some incomplete sentences**.**
summaries-gold/	• This directory contains human composed summaries (aka reference or model summaries) used as a gold standard for the Opinosis summarization paper. Each file contains a short summary written based on the topic based sentences. • On the average there are about 4 human composed summaries per topic.

	• The name of each child directory corresponds to a topic (as used in the topics folder). • The process of obtaining these summaries is described in the Opinosis paper.
scripts/	• This folder contains some helper perl scripts that could be useful for your evaluation or summarization tasks. • All usage instructions can be found within the scripts itself. Here is a summary of the scripts provided: **MEAD Related** The main point to note here is that you have to make sure that you set MEAD_HOME as follows for the scripts to work (assuming you use a linux system): `export MEAD_HOME=/<absolute path to mead directory>/.` • doc2mead.pl – Converts a set of text files into the docsent format by invoking text2cluster on each summarization task. You have to make sure that each summarization task has its own folder. For example, using this dataset, since there are 51 topics (51 summarization tasks), the text files are stored in 51 different folders. You need to be in the root directory of the task folders to run the script. • mead2summary.pl – This script runs the MEAD summarizer on the docsent files created. You can easily change the summarization parameters within this script. **ROUGE Related** • rouge2csv.pl - This script helps in interpreting ROUGE scores. If you need Instructions on how to set-up ROUGE for evaluation of your summarization tasks go here. Assuming you have piped all your ROUGE results to a file, this tool will collect all rouge scores into separate CSV files depending on the n-grams used. For example, all ROUGE-1 scores will be collected into a ROUGE-1.csv file, similarly all ROUGE-2 scores will be in a ROUGE-2.csv. The precision, recall and f-scores will be comma separated. This will allow you to easily visualize your results in Excel or OpenOffice. If you have ROUGE scores of identical runs (usually happens when you use Jackknifing), the scores will be averaged. For more information go here. • prepare4rouge.pl – This script prepares for ROUGE evaluation using a jackknifing procedure. If you do not want jackknifing, you may have to modify this script a little. The script basically takes in your baseline summaries, gold standard or reference summaries and system summaries to generate all files needed to evaluate using ROUGE (in the format that ROUGE understands). Example of input and output of this script is found in examples/ prepare4rouge. You need to replace a few variables in this file and it is clearly explained in the preamble of the script. For basic information on the usage of ROUGE you can go through my tutorial.
examples/	• All examples are placed in this folder.