==========

UNT Computer Science Short Answer Dataset
Version 2.0
April 22nd, 2011

==========

Rada Mihalcea, Michael Mohler

Language and Information Technologies
University of North Texas

rada@cs.unt.edu
mgm0038@unt.edu

CONTENTS

1. Introduction

This README v2.0 (April, 2011) for the automatic short-answer grading data set comes from the archive hosted at the following URL http://lit.csci.unt.edu/index.php/Downloads

2. Data Set

2.a. Data Annotation

This data set consists of ten assignments with between four and seven questions each AND two exams

with ten questions each. These assignments/exams were assigned to an introductory computer science class at the University of North Texas. The data is in plaintext format. Each assignment includes the question, instructor answer, and set of student answers with the average grades of two annotators included. Both annotators were asked to grade for correctness on an integer scale from 0 to 5. However, for the exams, the annotators provided a grade on a 0 to 10 scale. These scores were normalized to the 0..5 scale before being averaged.

For privacy reasons, no actual student identifiers are used in this corpus.

2.b. Folder Structure

For notational purposes, when we describe a file as "x.y.z" or "x.y", "x.y" indicates the question (generally x is the (1-based) assignment number, y is the (1-based) question number within that assignment) and z represents the z-th student answer (0-based) associated with the question.

The following data files are included:

File	Description
data/docs/files	A list of all files (in x.y format) in the dataset. Note that any lines beginning with a pound sign (#) are ignored in all of the authors' published work. Files that are being ignored are generally not short answer style questions, but rather selection/ordering style questions.
data/annotations/x.y.z	This set of files indicates the node-level alignments. All lines after ";Edge" can be safely ignored. The annotation format is [label] [instructor-node] [student-node] where label can be one of the following: 0 - indicates no match 1 - indicates that the entities could be matched ignoring context 2 - indicates that these entities would be matched if the two answers were aligned. Annotations 1 and 2 are combined in this work.

File	Description
data/scores/x.y/ave	These files each contain a list of grades assigned for question x.y -- one student grade per line.
data/scores/x.y/other	The grades actually given by grader1 (the TA associated with the class)
data/scores/x.y/me	The grades actually given by grader2 (Michael Mohler). Note that assignments 11 and 12 were graded on a 10 point scale (0..10). The average (reported in data/scores/x.y/ave) is normalized to be on the 0..5 scale, but these two files are raw and may contain values above 5 for these assignments.
data/raw/x.y	A list of raw student answers sentences associated with the question x.y -- one student answer per line.
data/raw/questions	All questions in a raw/uncleaned form
data/raw/answers	All instructor answers in a raw/uncleaned form.
data/sent/x.y	A list of cleaned student answers sentences associated with the question x.y -- one student answer per line.
data/sent/questions	All questions in sentence form
data/sent/answers	All instructor answers in sentence form. Sentences are separated with <STOP>. Sentence chunking was done using LingPipe toolkit.
data/parses/stanford.trip/x.y	A list of parses associated with the question x.y -- one student answer per line.
data/parses/stanford.trip/questions	All questions parsed
data/parses/stanford.trip/answers	All instructor answers parsed. • Sentences are separated with <STOP> • Dependency triples are separated with ||| • Parsing was performed using the Stanford Dependency Parser with some postprocessing

Formatted and styled by Nazmul Kazi

3. Feedback

For further questions or inquiries about this data set, you can contact: Michael Mohler (mgm0038@unt.edu) or Rada Mihalcea (rada@cs.unt.edu).

4. Citation Info

If you use this data set, please cite:

```
@InProceedings{mohler2011learning,
  author    = {Michael Mohler and Razvan Bunescu and Rada Mihalcea},
  title     = {Learning to Grade Short Answer Questions using Semantic
               Similarity Measures and Dependency Graph Alignments},
  booktitle = {Proceedings of the 49th Annual Meeting of the Association
               of Computational Linguistics: Human Language Technologies
               (ACL HLT 2011)},
  address   = {Portland, Oregon},
  year      = {2011}
}
```

5. Acknowledgments

This work was partially supported by a National Science Foundation grant #IIS-0747340. Any opinions, findings, conclusions or recommendations expressed above are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Formatted and styled by Nazmul Kazi