CS-5340/6340 Project Description, Fall 2015

The NLP class project will be to design and build a question answering system. For this project, we will use stories that were collected from the Canadian Broadcasting Corporation web page for kids. The CBC published five current events stories a week for over two years that targeted elementary and middle school students. Figure 1 shows a sample story.

HEADLINE: An Arctic Struggle
DATE: June 29, 1999
STORYID: 1999-W27-2

TEXT:

A group of 50 beluga whales is fighting to stay alive in an icy trap in the Canadian Arctic near Ellesmere Island.

An unexpected freeze has left dozens of the whales trapped in a sea of ice, with one small hole as their only window for air. The open sea is 20 kilometres away.

About 20 of them have already died, despite the best efforts of wildlife officials. The reason: polar bears. With the whales swarming their only air hole, they've become easy prey for the bears.

The polar bears are hunting the belugas at will. Even in the best of conditions, belugas can only spend about 10 minutes underwater. When they come up, the bears jump onto the whales and tear chunks out of them.

Unless the ice breaks up soon, the belugas' chances of survival are grim.

Beluga whales live only in the arctic and subarctic. They live in the Arctic Ocean and its adjoining seas, including the Sea of Okhotsk, the Bering Sea, the Gulf of Alaska, the Beaufort Sea, Baffin Bay, Hudson Bay, and the Gulf of St. Lawrence.

Figure 1: A story

Two people at the MITRE Corporation created questions and an answer key for each story. (One of these people had professional experience writing questions for reading comprehension exams.) Figure 2 shows the answer key for the story above. Each question has a unique *QuestionID* which is the *StoryID* followed by a dash and question number. For example, "1999-W27-2-1" means that this is question #1 pertaining to story "1999-W27-2". The questions have difficulty ratings assigned to them, as judged by the person who created the question. For this project, all of the questions will have a difficulty rating of "easy" or "moderate". Note that the QuestionID numbers may not be consecutive because questions that had a high difficulty were removed.

In some cases, the answer key allows for several acceptable answers (e.g., "Toronto, Ontario" vs. "Toronto"), paraphrases (e.g., "Human Immunodeficiency Virus" vs. "HIV"), varying amounts of information (e.g., "he died" vs. "he died in his sleep of natural causes"), or occasionally different interpretations of the question (e.g., "Where did the boys learn how to survive a storm?" "camping tips from a friend" vs. "their backyard"). When more than one answer is acceptable, the acceptable answers are separated by a vertical bar (|).

QuestionID: 1999-W27-2-1
Question: Where in the Canadian Arctic are the 50 beluga whales trapped?
Answer: near Ellesmere Island
Difficulty: easy

QuestionID: 1999-W27-2-2
Question: Why have the whales become trapped?
Answer: an unexpected freeze
Difficulty: moderate

QuestionID: 1999-W27-2-3
Question: Through what are the whales breathing?
Answer: one small hole in the ice | one small hole
Difficulty: moderate

QuestionID: 1999-W27-2-4
Question: Who is trying to save the whales?
Answer: wildlife officials
Difficulty: moderate

QuestionID: 1999-W27-2-5
Question: Even in the best of conditions, how long can beluga whales stay under water?
Answer: only about 10 minutes | about 10 minutes
Difficulty: moderate

QuestionID: 1999-W27-2-6
Question: How are the bears killing the whales?
Answer: the bears jump onto the whales and tear chunks out of them
Difficulty: moderate

Figure 2: The answer key for story 1999-W27-2

Judging answers is subjective, so you may occasionally disagree with MITRE's answers. But people will never completely agree on this kind of thing, and it is necessary to choose some set of answers for evaluation purposes, so we will stick with MITRE's judgements as the gold standard.

The Task: Your team must build a question answering (Q/A) system that can process a story and a list of questions, and produce an answer for each question. Each team will consist of two people, except for special cases approved by the instructor. Each team's system **must** conform to the following input and output specifications, but other than that you can design your system however you want!

The Input

Your Q/A system should accept a single input file as a command-line argument. We should be able to run your program like this:

```
qa <inputfile>
```

The first line of the input file will be a directory path. Each subsequent line in the file will be a StoryID. Your Q/A system should then assume that for each StoryID, the directory contains a story file named StoryID.story (e.g., "1999-W02-5.story") and a question file named StoryID.questions (e.g., "1999-W02-5.questions"). Your Q/A system should produce an answer for each question in the question file, based on the corresponding story file. A sample input file is shown below.

```
/home/cs5340/project/developset/
1999-W02-5
1999-W03-5
1999-W04-5
1999-W05-4
1999-W05-5
1999-W06-5
1999-W07-5
1999-W08-1
```

Each story file will be formatted like Figure 1. Each story will include a Headline, Date, and StoryID line, followed by the text of the story.

Each question file will contain 3 lines for each question indicating the QuestionID, the question itself, and a difficulty rating. The question file will be formatted like Figure 2, except that there will not be an answer line. For example, it would look like this:

```
QuestionID: 1999-W27-2-1
Question: Where in the Canadian Arctic are the 50 beluga whales trapped?
Difficulty: easy

QuestionID: 1999-W27-2-2
Question: Why have the whales become trapped?
Difficulty: moderate
```

The Output

Your Q/A system should produce a single *Response File*, printed to standard output, which contains the answers that your system finds for all of the stories and questions in the input file. The output of your system should be formatted as follows:

```
<QuestionID>
<Answer>

<QuestionID>
<Answer>

...
```

For example, your output should look like this:

```
QuestionID: 1999-W02-5-1
Answer: Canada

QuestionID: 1999-W02-5-2
Answer: Betty Jean Aucoin

QuestionID: 1999-W02-5-3
Answer: a fitness club

QuestionID: 1999-W03-5-4
Answer: 5 feet, 2 inches

QuestionID: 1999-W03-5-5
Answer:

QuestionID: 1999-W03-5-6
Answer: 502

QuestionID: 1999-W02-5-7
Answer: $135

QuestionID: 1999-W03-5-2
Answer: Edmonton

QuestionID: 1999-W03-5-3
Answer: forward

...
```

IMPORTANT: Your response file should have a QuestionID and Answer for **every question** in **every story** specified by the input file, in **exactly the same order**. Also, be sure to print each answer on a single line. If your Q/A system can't find an answer to a question (or your system chooses not to answer a question), then leave the answer blank (as done for Question 1999-W03-5-5 above).

The Data Sets

You will be given three sets of data at different points in the project.

Development Set: 73 stories and answer keys

Test Set #1: 39 stories and answer keys

Test Set #2: 39 stories and answer keys

Project Phases

The project will involve three phases:

Development Phase: A **Development Set** is available on our class web page for you to use when creating your Q/A system. You may use these stories and the answer keys in any way that you wish. The development data set can also be found here:

/home/cs5340/project/developset/

In addition, we will give you the scoring program that we will use to evaluate your Q/A system. You can use this scoring program to assess the performance of your system yourself as you experiment with different ideas. The scoring program is available on our web page, or you can find it here:

/home/cs5340/project/score-answers.pl

The arguments that it takes are described at the beginning of the file.

Dry Run Evaluation: On Thursday November 12, there will be a dry run evaluation of everyone's Q/A systems. Each team must hand in the code for their Q/A system and we will evaluate each system on a brand new data set, called **Test Set #1**. The performance of your Q/A system during the dry run evaluation will account for 10% of your project grade. There will be four possible grades:

Excellent (100/100): System achieves $\geq 25\%$ F measure score.

Very Good (90/100): System achieves $< 25\%$ but $\geq 10\%$ F measure score.

Satisfactory (80/100): System achieves $< 10\%$ but $> 0\%$ F measure score, has correctly formatted output, and makes good faith attempt to answer questions based on the story.

Fail (0/100): No system submitted, system does not produce correctly formatted output, or system does not exhibit a "good faith attempt" to answer questions based on the corresponding story (e.g., answers every question with "the"!). Your system does **not** have to answer every question, but it should attempt to answer *some* questions.

Once the dry run is over, we will release Test Set #1 so that you can improve the performance of your system on those stories.

Final Evaluation: On December 1, each team will hand in the final code for their Q/A system. We will run your Q/A system on **both** Test Set #1 and Test Set #2. Your final project grade will be based on the performance of your Q/A system on both test sets.

The purpose of evaluating your system on both test sets is to balance specificity with generality. You will have 3 weeks to try to get your Q/A system to perform well on Test Set #1. Hopefully, everyone will be able to do fairly well on that test set. Test Set #2 will be a blind test set that no one will see until the final evaluation. A system that uses general techniques should work just as well on Test Set #2 as Test Set #1. But a system that has lots of hacks and tweaks based on Test Set #1 probably will perform poorly on Test Set #2.

WARNING: You will be given the answer keys for Test Set #1, but your Q/A system is not allowed to use them to answer questions! For example, you can **not** just look up the answer to each question, or use the answer keys as training data for a machine learning algorithm. Your system must answer each question using general methods, and you must use **exactly the same system on both test sets**. The answer keys are being distributed only to show you what the correct answers are, and to allow you to score your system's performance yourself.

Evaluation

The performance of each Q/A system will be evaluated using the F-measure statistic, which combines recall and precision in a single evaluation metric. Since Q/A systems often produce answers that are partially but not fully correct, we will score each answer by computing the *Word Overlap* between the system's answer and the strings in the answer key. Given an answer string from the answer key, your system's response will be scored as:

Recall (R): the number of correct words generated by your system divided by the total number of words in the answer string.

Precision (P): the number of correct words generated by your system divided by the total number of words generated by your system.

F-measure: $\mathbf{F(R, P)} = \frac{2 \times P \times R}{P + R}$
This formula tries to find a good balance between recall and precision. (It is the harmonic mean of recall and precision.) *The final performance of each system will be based on its F-measure score.*

As an example, suppose your system produces the answer "Elvis is great" and the correct answer string was "Elvis Presley". Your system's answer would get a recall score of 1/2 (because it found "Elvis" but not "Presley"), a precision score of 1/3 (because 1 of the 3 words that it generated is correct), and an F-measure score of .40.

Two important things to make a note of:

- This scoring measure is not perfect! You will sometimes receive partial credit for answers that look nothing like the true answer (e.g., they both contain "of" but all other words are different). And you may sometimes get a low score for an answer that seems just fine (e.g., it contains additional words that are related to the true answer, but these extra words aren't

in the answer key). Word order is also not taken into account, so if your system produces all the correct answer words, but in the wrong order, it doesn't matter – your system will get full credit! Automatically grading answers is a tricky business. This metric, while not perfect, is meant to be generous and give your system as much partial credit as possible.

- The answer key often contains multiple answer strings that are acceptable for a question. Your system will be given a score based on the answer string that mostly closely matches your system's answer.

Schedule

October 9: Fill in the Team Request Form on the class web site and submit it via electronic handin: *handin cs5340 team-info team-form.txt*

November 12: Dry run evaluation on Test Set #1. Test Set #1 will be released after the dry run.

December 1: Final evaluation on Test Set #1 and Test Set #2.

December 7, 9: Project presentations.

December 11: Project summaries due.

By November 12, we expect each team to have a working Q/A system! It might not work well and may still be missing components that you plan to incorporate, but it should be able to process a story and make a good faith attempt to produce some answers.

Details on the presentations and project summaries will be forthcoming.

Grading

Each project will be graded according to the following criteria:

- 10% of the grade will be based on the performance of your Q/A system on Test Set #1 during the dry run evaluation.
- 30% of the grade will be based on the performance of your Q/A system on Test Set #1 during the final evaluation.
- 35% of the grade will be based on the performance of your Q/A system on Test Set #2 during the final evaluation.
- 25% of the grade will be based on your project presentation, summary, and the originality and ambitiousness of your system's design.

To compute the final grade for the project, each Q/A system will be ranked on the last 3 criteria listed above, relative to all of the other Q/A systems. We will then compute an "average" ranking for each system, which will be used to assign final grades. For example, if your system ranks 2nd

on Test Set #1 during the final evaluation, 3rd on Test Set #2 during the final evaluation, and 5th on the presentation/report/originality component, then the average ranking would be (2+3+5)/3 = 3.33

IMPORTANT: I will also ask each team to document the specific contributions of each team member to the team's final Q/A system. This will allow me to adjust individual grades in case some people work a lot harder than others. If all teammates contribute roughly equally to the project (as I expect in most cases), then they will all get the same project grade. But if one teammate contributes very little to the project, then they will get a lower grade than the person who put in substantially more time and effort.

The final grading will therefore be based on how well your system does relative to other team's systems, but it is not the case that the highest ranked system will get an automatic 'A' or that the lowest ranked system will get an automatic 'E'. If every team produces a system that works well, then I will be happy to give everyone an 'A' on the project! If, at the other extreme, no one generates a system that works at all, then I would have to give every team a failing grade. I hope that the competitive spirit will energize everyone to work hard and produce interesting and effective Q/A systems so that I can give many teams a high grade!

CAVEAT AND ENCOURAGEMENT

Building an effective question answering system is hard! Question answering is not a solved problem in NLP, and the performance of the very best research systems is still mediocre. So you shouldn't expect your Q/A systems to produce super-high scores! Just try to design a Q/A system as best you can and try to develop a Q/A system that can perform better than the Q/A systems developed by the other teams in the class.

I encourage everyone to have fun with this project and experiment with lots of ideas. Creativity is appreciated! I chose this project because question answering is an important NLP application. Right now, it is also a hot research area, but hopefully in the next 10 years the technology will be good enough to start incorporating Q/A into real products. This project will give you exposure to a cutting edge research area, understanding of an important application area for NLP, the experience of building a real NLP system, and the opportunity to explore your creative side!