Project Phase II

Team LeNguyen: Trung Le, Michelle Nguyen

1 Baseline II

For our second baseline, we tried to make our baseline I scores improve by using new techniques for extraction for the more challenging slots which will be described below.

The techniques used for *"Story," "ID," "Event,"* and *"Country"* did not change due to the fact that these slots are pretty self-explanatory. *"Story"* was given by the filename, *"ID"* is always 1 for every article, and *"Event"* is always *"outbreak"* since that is the domain we are dealing with. *"Country"* was found using a compiled list of countries found online and if any countries from the list were found in the article, it was listed in the *"Country"* slot (this includes abbreviations of country names as well).

For the *"Date"* slot, we discovered that it could be found from the filename. In most cases, the first four numbers represents the year of the article, then the next two numbers represents the month, and the following two numbers represents the day.

For the *"Status"* and *"Containment"* slots, we used a Bag-of-Words model to generate feature vectors for each file and then used liblinear in order to make predictions on the test articles. For the Bag-of-Words word mappings, we assigned each possible status and containment a value. Then for each unique word in all of the files, we also assigned it a value. Afterwards, for each file, we printed the value pertaining to the slot's answer, then next to the answer a word mapping. For each word that appeared in the file, we printed the value mapped to the word and then a colon followed by its frequency count (e.g 2 1332:8 2751:12). Afterwards, we passed the vector files into liblinear to perform classification. Liblinear is an external machine learning library used for linear classification. It takes in all the vectors for each file in the development set in the format described above and performs a prediction.

The *"Disease"* and *"Victims"* slots were discovered using the external Stanford NER tagger. We trained the NER classifier on the development set by labeling the correct answers in the development articles as either "DIS" (for disease) or "VIC" (for victim). Other words are labeled according to their NER tag. Afterwards, we used the trained classifier in order to label these previously found diseases and victims in the test set.

Our system works from the command line. For the second baseline system, we made it more interactive by allowing users to initially pick from three options. Users can choose to pass in the test folder, pass in a single file of their choice, or to edit a file and pass it in. The first two options are the same as last time with each file passed in being returned the generated template, answer template, ans scores for each template. However, the third option is newly-added. When users edit an original file, scores for the edited file will not be shown to them, due to lack of an answer template for the edited file, however, users will be shown the generated template for their edited file as well as the template for the original file to serve as comparison.

2 External Resources

In order to perform Named Entity Recognizer (NER) tagging, we used the external tagger tool from Stanford Core NLP. The tool and information about it can be found at the URL:

https://nlp.stanford.edu/software/tagger.shtml

The online source used to compile the list of countries is found here:

https://www.state.gov/misc/list/index.htm

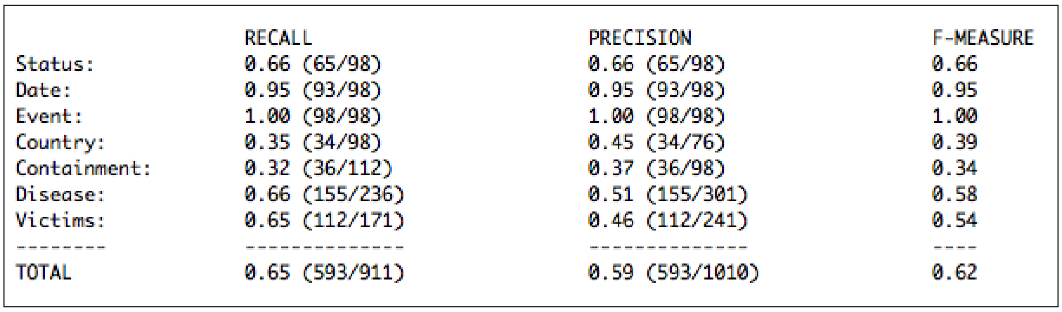
The liblinear tool used to in conjunction with the Bag-of-Words model can be found here:

https://www.csie.ntu.edu.tw/ cjlin/liblinear/

3 Evaluation

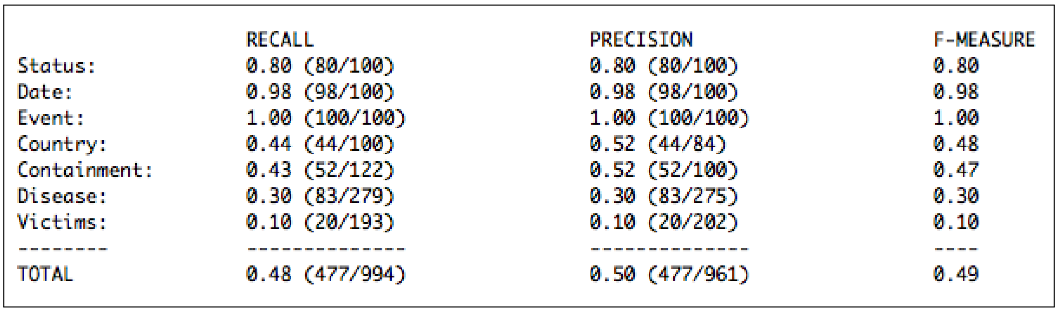
The disease outbreak dataset contains 198 original text files as well as 198 annotated files with templates for each corresponding document. 98 of the document and annotated files were used as the development set and were examined to test and tweak the system and the rules. The 100 remaining original and annotated files were used a blind test set and were only used to evaluate the final baseline IIII system, not in development. Both the development and test sets contain roughly an equal amount of articles from the years 2000-2004.

In order to evaluate our program, we used the same scoring program from Baseline I that measures recall, precision, and F-scores for each slot using the provided templates in the dataset. For answer templates that contained an answer with a conjunction of multiple answers, the scoring program considered the answer correct if the generated template correctly got one of the answers from the conjunction. The average performance measures our system got for the development set were as follows:



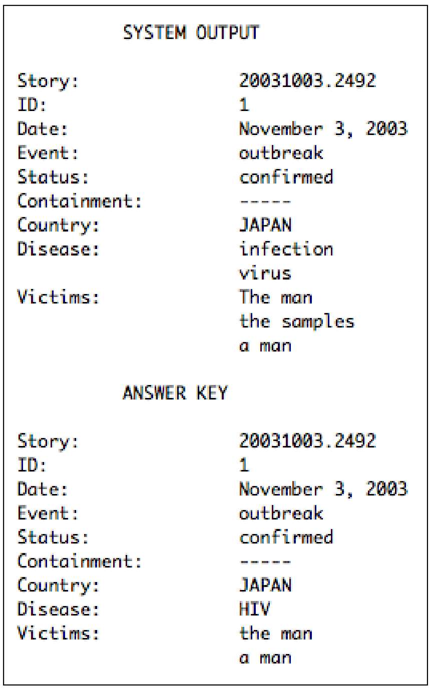
```
                     RECALL                PRECISION              F-MEASURE
Status:              0.66 (65/98)          0.66 (65/98)           0.66
Date:                0.95 (93/98)          0.95 (93/98)           0.95
Event:               1.00 (98/98)          1.00 (98/98)           1.00
Country:             0.35 (34/98)          0.45 (34/76)           0.39
Containment:         0.32 (36/112)         0.37 (36/98)           0.34
Disease:             0.66 (155/236)        0.51 (155/301)         0.58
Victims:             0.65 (112/171)        0.46 (112/241)         0.54
--------             --------------        --------------         ----
TOTAL                0.65 (593/911)        0.59 (593/1010)        0.62
```

The average performance measures our system got for the test set were as follows:



```
                     RECALL                PRECISION              F-MEASURE
Status:              0.80 (80/100)         0.80 (80/100)          0.80
Date:                0.98 (98/100)         0.98 (98/100)          0.98
Event:               1.00 (100/100)        1.00 (100/100)         1.00
Country:             0.44 (44/100)         0.52 (44/84)           0.48
Containment:         0.43 (52/122)         0.52 (52/100)          0.47
Disease:             0.30 (83/279)         0.30 (83/275)          0.30
Victims:             0.10 (20/193)         0.10 (20/202)          0.10
--------             --------------        --------------         ----
TOTAL                0.48 (477/994)        0.50 (477/961)         0.49
```

As we can see, the predictions for diseases and victims performed a lot better on the development set than the test set. This is because the NER classifier uses previously seen diseases and victims in order to predict future diseases and victims. Diseases from the development set were sometimes repeated in the test set but only about a quarter of the time. The victims encountered in the development set were hardly ever the same victim again in the test set, hence the low values for recall and precision for the "Victim" slot in the test set. The only time victims were properly detected in the test set was when the victim were generic victims such as "the man" or "[number of] cases" as can be seen below:

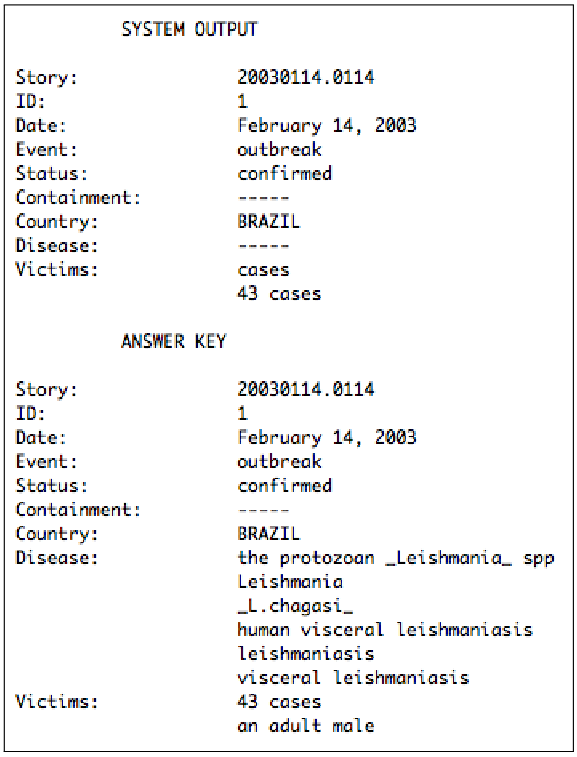


```
          SYSTEM OUTPUT

Story:                   20031003.2492
ID:                      1
Date:                    November 3, 2003
Event:                   outbreak
Status:                  confirmed
Containment:             -----
Country:                 JAPAN
Disease:                 infection
                         virus
Victims:                 The man
                         the samples
                         a man

          ANSWER KEY

Story:                   20031003.2492
ID:                      1
Date:                    November 3, 2003
Event:                   outbreak
Status:                  confirmed
Containment:             -----
Country:                 JAPAN
Disease:                 HIV
Victims:                 the man
                         a man
```



```
          SYSTEM OUTPUT

Story:                   20030114.0114
ID:                      1
Date:                    February 14, 2003
Event:                   outbreak
Status:                  confirmed
Containment:             -----
Country:                 BRAZIL
Disease:                 -----
Victims:                 cases
                         43 cases

          ANSWER KEY

Story:                   20030114.0114
ID:                      1
Date:                    February 14, 2003
Event:                   outbreak
Status:                  confirmed
Containment:             -----
Country:                 BRAZIL
Disease:                 the protozoan _Leishmania_ spp
                         Leishmania
                         _L.chagasi_
                         human visceral leishmaniasis
                         leishmaniasis
                         visceral leishmaniasis
Victims:                 43 cases
                         an adult male
```

As you can see, for the story on the left, our system got a man and the man because those were victims seen before in previous development articles. It incorrectly labeled Evaluations of our scores and program performance can be found in the "Evaluations.pdf" file included with the submission.

4 Member Contributions

Trung Le used Stanford Core NLP in order to train the NER classifier for the Disease and Victim slots. He also used generated word vectors with liblinear in order to train the classifier for the Status and Containment slots. Trung also discovered how to determine the date for each file and implemented this labeling scheme as well.

Michelle Nguyen made changes to the interface by making it more interactive. She also wrote the class that generated the bag-of-word feature vectors for each file in order to be used for the liblinear classification.