Project Phase I

Team LeNguyen:
Trung Le
Michelle Nguyen

1 Baseline I

Our system is a rule-based extraction system that takes in articles pertaining to disease outbreaks as input and returns filled-out slots about each article in a template format. The slots to be filled out include: "Story," "ID," "Date", "Event", "Status", "Containment," "Country," "Disease," and "Victim". The difficulty of the problem depends on the slots themselves. Some slots are more trivial to answer than others.

For the *"Story"* slot, the system returns the name of the file in question. The ID for all of the articles is the number 1, therefore the system always returns 1 in the *"ID"* slot. For the *"Event"* slot, the system always returns "outbreak" since we are analyzing the disease outbreak domain.

For the *"Status"* slot, we checked if the article contained the words "confirmed," "possible," or "suspected". If so, the status of the article was labeled as the word found in the file. If more than one of those words appeared, a hierarchical labeling scheme was adopted with "confirmed" having highest priority, "suspected" with second priority, and "possible" with the lowest priority. Articles that contained neither of these three options were automatically assigned "confirmed" due to it being the most frequent of the three labels.

In order to fill out the *"Country"* slot, we compiled a list of countries using an online source (see section 2). Afterwards, any countries from the list that were found in the articles were labeled in the "Country" slot. Anytime abbreviations for the United States or the United Kingdom appeared in articles, the system detected these as well and labeled the respective countries based on their acronyms.

The *"Disease"* and *"Victims"* slots were discovered using manually generated patterns and an external NER/POS tagger. Context files were made with both text and answer files to find instances from the text where each answer was mentioned. Afterwards, these context sentences aided in the discovery of patterns often used when describing diseases or victims mentioned in the articles. The patterns were then applied to the text with one or more of the words from the pattern acting as a trigger. When a trigger is reached, the system looks to the left or right of the trigger (depending on the pattern) and returns the nearest noun phrase as the answer.

The *"Date"* slot was left unanswered by the system because dates were not always in the original articles, however, the answer templates contained dates with no apparent solution to be extracted from the text, therefore we chose not to label any articles with dates. In addition for this baseline system, the *"Containment"* slot was filled in using a majority vote.

The system works from the command line. Users can choose to pass in a folder containing files which each represent disease outbreak texts or they can pass in a single file representing one disease outbreak article using the paths of either the folder or the file. A flag "-s" will be passed in as an argument following the folder or file path, which denotes that the user is passing in a *s*ingle file. Otherwise, if the flag is omitted, the system will search for a folder of files as input to the system instead.

For each disease outbreak article that is passed into the system, the system will print to console the answer key, the generated template, and the scores that the system received for each file.

2 External Resources

In order to perform Parts-of-Speech (POS) and 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:

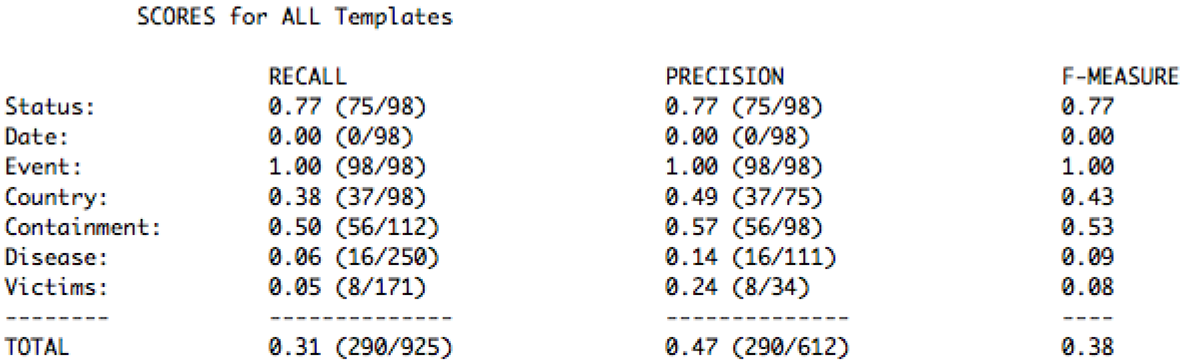
https://www.state.gov/misc/list/index.htm

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 1 system, not in development.

In order to evaluate our program, we wrote a scoring program that measures recall, precision, and F-score for each slot in each article using the provided templates in the dataset. In addition, average performance measure scores for each slot as well as total average scores are calculated after all of the passed in files have been evaluated. For the victim and disease slots, if multiple answers exist in the answer template and the system got at least one of the answers correctly, this was considered correct and was added to the recall and precision measures. Articles with multiple templates generated for multiple countries were only considered correct if individual templates were generated for each country detected.

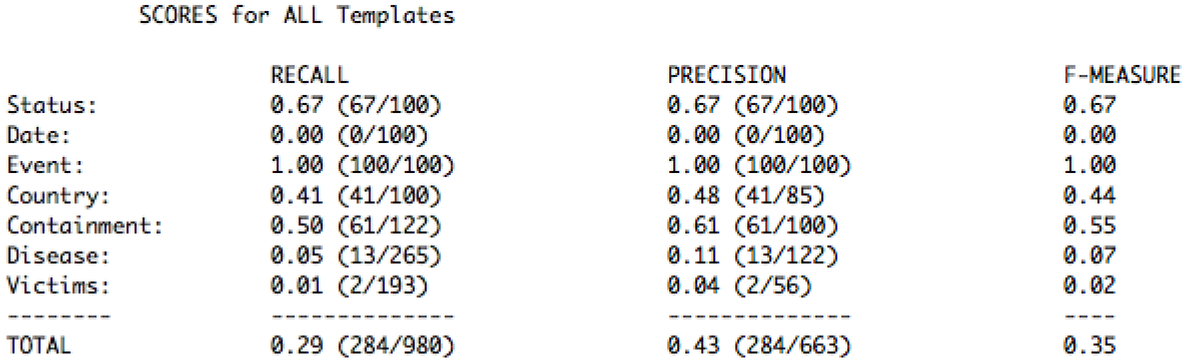
The average performance measures our system got for the development are shown:



```
          SCORES for ALL Templates

                   RECALL                        PRECISION                     F-MEASURE
Status:            0.77 (75/98)                  0.77 (75/98)                  0.77
Date:              0.00 (0/98)                   0.00 (0/98)                   0.00
Event:             1.00 (98/98)                  1.00 (98/98)                  1.00
Country:           0.38 (37/98)                  0.49 (37/75)                  0.43
Containment:       0.50 (56/112)                 0.57 (56/98)                  0.53
Disease:           0.06 (16/250)                 0.14 (16/111)                 0.09
Victims:           0.05 (8/171)                  0.24 (8/34)                   0.08
--------           --------------                --------------                ----
TOTAL              0.31 (290/925)                0.47 (290/612)                0.38
```

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



```
          SCORES for ALL Templates

                   RECALL                        PRECISION                     F-MEASURE
Status:            0.67 (67/100)                 0.67 (67/100)                 0.67
Date:              0.00 (0/100)                  0.00 (0/100)                  0.00
Event:             1.00 (100/100)                1.00 (100/100)                1.00
Country:           0.41 (41/100)                 0.48 (41/85)                  0.44
Containment:       0.50 (61/122)                 0.61 (61/100)                 0.55
Disease:           0.05 (13/265)                 0.11 (13/122)                 0.07
Victims:           0.01 (2/193)                  0.04 (2/56)                   0.02
--------           --------------                --------------                ----
TOTAL              0.29 (284/980)                0.43 (284/663)                0.35
```

Evaluations of our scores and program performance can be found in the “Evaluations.txt” file included with the submission.

4 Member Contributions

Trung Le wrote the scoring program described in section 3 that was used to evaluate our system. In addition, he implemented the extraction methods for the “Country” and “Status” slots. He also used our generated patterns and the Stanford Core NLP tagger to tag the sentences in the articles and find and return labeled answers in the templates.

Michelle Nguyen wrote the script that generated the “Context Files” described in section 1. In addition, she wrote the rules for the “Disease” and “Victims” slots, which were used to detect diseases and victims in the articles. She also wrote some rules for the “Containment” slot which at the moment has not been put to use yet.