Evaluation

Development Set Results

	Baseline I Recall	Baseline II Recall	Baseline I Precision	Baseline II Precision	Baseline I F-Score	Baseline II F-Score
Status	0.72 (71/98)	0.66 (65/98)	0.72 (71/98)	0.66 (65/98)	0.72	0.66
Date	0.00 (0/98)	0.95 (93/98)	0.00 (0/98)	0.95 (93/98)	0.00	0.95
Event	1.00 (98/98)	1.00 (98/98)	1.00 (98/98)	1.00 (98/98)	1.00	1.00
Country	0.35 (34/98)	0.35 (34/98)	0.45 (34/76)	0.45 (34/76)	0.39	0.39
Containment	0.56 (63/112)	0.32 (36/112)	0.64 (63/98)	0.37 (36/98)	0.60	0.34
Disease	0.06 (14/236)	0.66 (155/236)	0.14 (14/100)	0.51 (155/301)	0.08	0.58
Victims	0.02 (4/171)	0.65 (112/171)	0.11 (4/38)	0.46 (112/241)	0.04	0.54
TOTAL	0.31 (284/911)	0.65 (593/911)	0.47 (284/606)	0.59 (593/1010)	0.37	0.62

Test Set Results

	Baseline I Recall	Baseline II Recall	Baseline I Precision	Baseline II Precision	Baseline I F-Score	Baseline II F-Score
Status	0.71 (71/100)	0.80 (80/100)	0.71 (71/100)	0.80 (80/100)	0.71	0.80
Date	0.00 (0/100)	0.98 (98/100)	0.00 (0/100)	0.98 (98/100)	0.00	0.98
Event	1.00 (100/100)	1.00 (100/100)	1.00 (100/100)	1.00 (100/100)	1.00	1.00
Country	0.44 (44/100)	0.44 (44/100)	0.52 (44/84)	0.52 (44/84)	0.48	0.48
Containment	0.44 (54/122)	0.43 (52/122)	0.54 (54/100)	0.52 (52/100)	0.49	0.47
Disease	0.05 (15/279)	0.30 (83/279)	0.11 (15/133)	0.30 (83/275)	0.07	0.30
Victims	0.03 (6/193)	0.10 (20/193)	0.12 (6/52)	0.10 (20/202)	0.05	0.10
TOTAL	0.29 (290/994)	0.48 (477/994)	0.43 (290/669)	0.50 (477/961)	0.35	0.49

Event scores did change. It will always have a recall of 100% since it is always “outbreak”.

Country scores also have not changed since Phase I. Using the method of creating a list of all countries and returning a country that is mentioned in the article as the country of the disease outbreak produces around 35-52% recall and precision. This technique does not produce the highest scores because most of the articles mention more than one country when talking about origin of the disease or about other factors that relate to places. Just because a country is mentioned does not necessarily mean that is where the outbreak in question has taken place.

Date was implemented for this phase using the filename. Not all the dates in the answer templates match the date listed in the file name, hence why the F-scores for this slot were not always 1.0. However, since we did not implement finding the date in phase I, we have made a huge improvement in the F-score for the date slot —going from scores of 0.0 to nearly an average score of 1.0 in phase II.

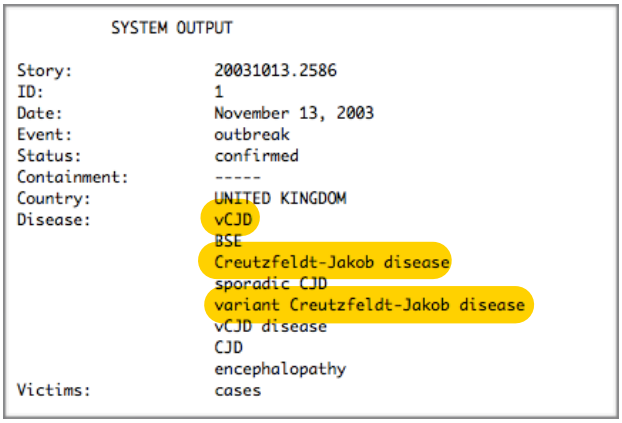
Status and **Containment** in Phase II were implemented using a Bag-of-Word model. In Phase I, we used a most frequent labeling scheme, therefore the status was always set to “confirmed”

since that was the most frequent label and the containment was always set to "-----" since that was the most common label for the containment slot. The results for these slots when comparing baseline I vs. baseline II scores and development vs. test set scores don't really have an explainable pattern. For the development set, we see the the F1-scores for status decreased from baseline I to baseline II, while the scores for containment dropped drastically. However, for the test set, the F1-scores for status increase going from baseline I to baseline II while the containment score dropped a little. The Bag-of-Words model doesn't perform that well for the containment slot because more than half of the articles did not mention a specific containment method. Therefore, the words used in each article doesn't really point to the containment. Maybe it would be better to classify the containment methods as "-----" by looking for a lack of words instead.

As for the status, it was hit-or-miss whether the Bag-of-Words model for classification was an improvement or worsening of F1-scores, therefore this tells us that it does not perform better than the most frequent labeling scheme in all cases and is not a good way to predict the statuses of the articles.

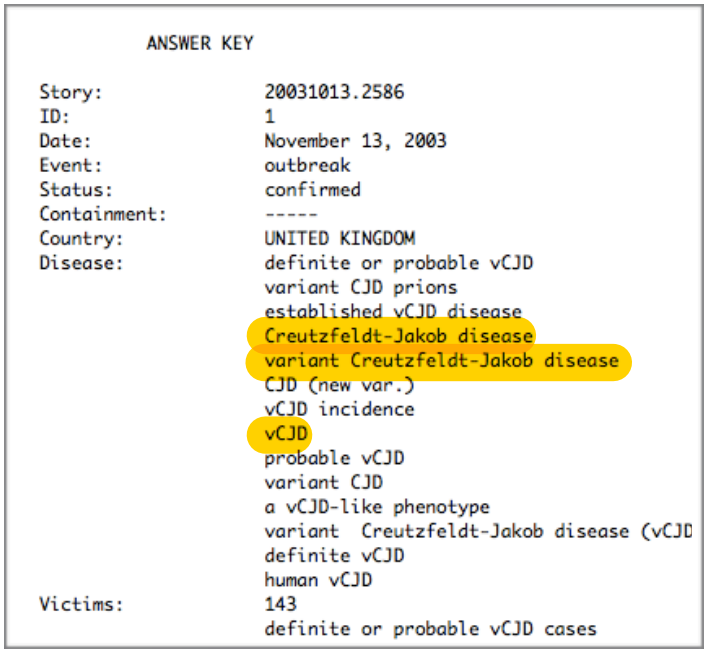
Disease and **Victims** increased dramatically going from phase I to phase II (more so for the development set than the test set). In phase I, we were using manually generated patterns and searching for these patterns in the articles to detect diseases and victims. This performed quite poorly once actually put to practice. For phase II, we used a NER tagger in training and tagged the answers from the answer keys in the original articles as "VIC" or "DIS". Afterwards, the NER classifier used the context of the diseases or victims to detect and label victims and diseases in the text. It is clear why the classifier performed better on the development set than the test set because the classifier was trained on the development set and had seen the diseases and victims in that context before.

For the test set, most of the time, a certain disease was detectable, however, it was very rare that all instances and different variations of the disease name were. Therefore, for stories where a disease was mentioned multiple times and in different ways, our system usually only returned 2-3 variations of the disease name. This would normally equate to high precision and low recall, however, the NER tagger seems to have tagged a "of", "cow" and "a" sometimes as diseases, which brought down the average precision as well. Some examples are below:



SYSTEM OUTPUT

Story:	20031013.2586
ID:	1
Date:	November 13, 2003
Event:	outbreak
Status:	confirmed
Containment:	-----
Country:	UNITED KINGDOM
Disease:	vCJD
	BSE
	Creutzfeldt-Jakob disease
	sporadic CJD
	variant Creutzfeldt-Jakob disease
	vCJD disease
	CJD
	encephalopathy
Victims:	cases



ANSWER KEY

Story:	20031013.2586
ID:	1
Date:	November 13, 2003
Event:	outbreak
Status:	confirmed
Containment:	-----
Country:	UNITED KINGDOM
Disease:	definite or probable vCJD
	variant CJD prions
	established vCJD disease
	Creutzfeldt-Jakob disease
	variant Creutzfeldt-Jakob disease
	CJD (new var.)
	vCJD incidence
	vCJD
	probable vCJD
	variant CJD
	a vCJD-like phenotype
	variant Creutzfeldt-Jakob disease (vCJD
	definite vCJD
	human vCJD
Victims:	143
	definite or probable vCJD cases

```
SYSTEM OUTPUT

Story:          20040511.1271
ID:             1
Date:           June 11, 2004
Event:          outbreak
Status:         confirmed
Containment:    -----
Country:        ISRAEL
Disease:        disease
                foot and mouth disease
                FMD
Victims:        -----
```

```
ANSWER KEY

Story:          20040511.1271
ID:             1
Date:           June 11, 2004
Event:          outbreak
Status:         confirmed
Containment:    vaccine
                quarantine
                inspection
Country:        ISRAEL
Disease:        Foot and mouth disease
                foot and mouth disease (FMD)
                the FMD situation
                FMD
Victims:        8 infected farms
                young fattening cattle and sheep
```

For the victims slot, only generic victims that were encountered in the development set could be identified. This is because victim names are so specific that the NER tagger had difficulties labelling the victim as a victim and instead tagged them as any other NP. Therefore the 0.10 F1-score for the test set was due to correctly labelling common victims such as “a woman” and “43 cases”:

```
SYSTEM OUTPUT

Story:          20030114.0114
ID:             1
Date:           February 14, 2003
Event:          outbreak
Status:         confirmed
Containment:    -----
Country:        BRAZIL
Disease:        -----
Victims:        cases
                43 cases
                3892 human 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
```

```
SYSTEM OUTPUT

Story:          20030808.1954
ID:             1
Date:           September 8, 2003
Event:          outbreak
Status:         confirmed
Containment:    -----
Country:        AUSTRALIA
Disease:        Variant Creutzfeldt-Jakob disease
                vCJD
                sporadic
                BSE
                sporadic CJD
                of
                cow disease
                CJD
Victims:        cases
                New Zealanders
                A Waikato man
                man
```

```
ANSWER KEY

Story:          20030808.1954
ID:             1
Date:           September 8, 2003
Event:          outbreak
Status:         suspected
Containment:    -----
Country:        NEW ZEALAND
Disease:        Variant Creutzfeldt-Jakob disease
                vCJD
                Variant CJD
                an "undiagnosed progressive  neurological disease"
                the [sporadic and hereditary] forms of CJD
                [as possibly  a case of variant] CJD
                CJD (new var.)
Victims:        the man
                A Waikato man
                The Waikato case
                The Waikato patient
```

Recall for victims was extremely low due to the system returning a lot of random nouns as seen below:

SYSTEM OUTPUT

Story:	20030310.0589
ID:	1
Date:	April 10, 2003
Event:	outbreak
Status:	confirmed
Containment:	-----
Country:	PARAGUAY
Disease:	dengue fever
	the virus
	Dengue serotype 2
	Dengue
	dengue virus
	The Dengue
	fever
	the dengue virus
	Dengue virus
	dengue
	virus
Victims:	at least 315 people
	cases
	000 cases
	imported cases
	3 cases

ANSWER KEY

Story:	20030310.0589
ID:	6
Date:	April 10, 2003
Event:	outbreak
Status:	confirmed
Containment:	pesticide
Country:	INDONESIA
Disease:	Chikungunya Outbreaks
	chikungunya
Victims:	99 cases

SYSTEM OUTPUT

Story:	20030925.2422
ID:	1
Date:	October 25, 2003
Event:	outbreak
Status:	confirmed
Containment:	-----
Country:	GEORGIA
Disease:	West Nile virus
	West Nile virus infection
	WNV infections
	WNV infection
	virus
	WNV
Victims:	2 human cases
	New York
	New
	cases
	human cases
	young people
	New Jersey
	seropositive
	The 18-year-old man
	New Mexico
	Connecticut
	6 positive
	4827 human cases

ANSWER KEY

Story:	20030925.2422
ID:	1
Date:	October 25, 2003
Event:	outbreak
Status:	confirmed
Containment:	-----
Country:	UNITED STATES
Disease:	West Nile virus
	WNV fever
	West Nile virus infection
	West Nile virus (WNV)
	WNV infection
	WNV
Victims:	unidentified animal species
	8406 dead birds
	603 sentinel chicken flocks
	dogs
	4827 human cases
	horses