





The ProMED-mail Coreference Corpus:  
Outbreak Detection Reports  
Annotated for Coreference Resolution


Scott L. DuVall, Jeffrey P. Ferraro,  
Ellen Riloff



ISDS 2009



Do you subscribe to  
ProMED-mail  
or a similar service?





How many emails do you  
receive in a week?



Would you be willing to  
receive more if you knew  
they were relevant to  
your interests?



**Thailand suspects  
new outbreak of bird  
flu**

Thailand suspects a new outbreak of bird flu at a farm in a central province where thousands of chickens have died, the deputy agriculture minister said on Tue 6 Jul 2004.

outbreak:

bird flu

affected locations:

central province in  
Thailand

suspected cases: 1000's

victims: chickens



natural language processing



## coreference resolution



## Types of Coreference

Acronyms (H1N1, nvCJD)

Pronouns (it, she)

Appositives (Ms. Shuchinova, the director)

Generic phrases (the disease, the official)

Descriptive phrases

(one of the worst outbreaks, swine flu)



coreference resolution



Thailand suspects  
new outbreak of bird  
flu

... was free of the  
disease following  
widespread outbreaks ..

...a new outbreak of bird  
flu at a farm in a central  
province ...

... where thousands have  
died ...

outbreak:

the disease

affected locations:

a farm in a central  
province

suspected cases: 1000's

victims: ?



## Coreference Resolution



## Available Training Sets



MUC





# The ProMED-mail Coreference Corpus

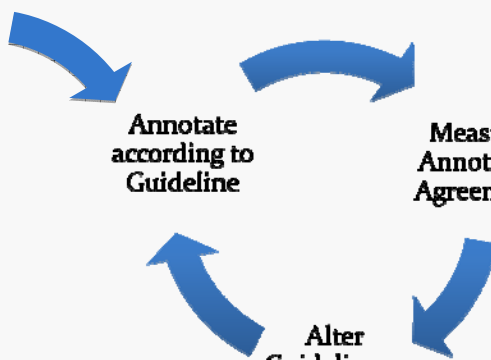


Original  
Guideline  
(MUC 7)

Annotate  
according to  
Guideline

Measure  
Annotator  
Agreement

Alter  
Guideline as  
appropriate





# Annotation Tool

Annotation Viewer

1927\_ferraro (33)

West Nile virus positive crow found in North Carolina

the  
the  
far  
inf  
40  
We  
the  
car  
pen  
on  
Us  
off  
per  
the  
ten  
car  
car

WASHINGTON: On Sat 20 Oct 2000 U.S. gov  
they had found **West Nile virus** in a dead crow  
farthest south **the virus** has been found. The l  
infected crow was found in Chatham County, r  
40 miles southwest of Raleigh.

**West Nile virus** naturally infects birds and is s  
they bite people, **it** can create a mild flu-like ill  
can go on to cause encephalitis. The enceph

COREF - ID: 0 [West Nile virus]  
COREF - ID: 2 [West Nile virus positive]  
COREF - ID: 4 [North Carolina]  
COREF - ID: 28 [U.S.]  
COREF - ID: 30 [U.S. government resea  
COREF - ID: 31 REF: 30 [they]

next ID: 35  
show all mentions  
 sort by chain

Delete

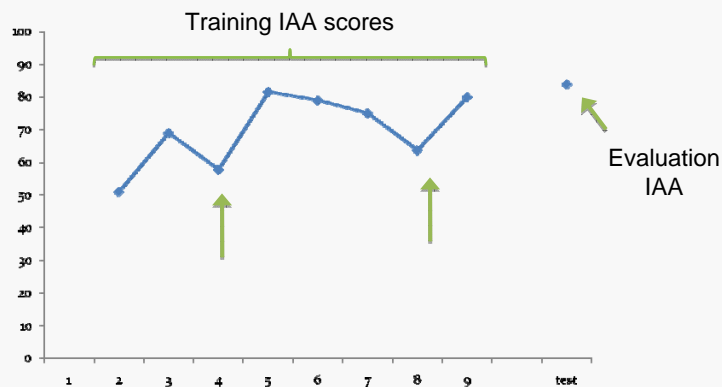
Annotation: string match

Start: 203  
End: 218  
ID: 1  
REF: 0

Prev C:\javaltool\annotator\annotate\_mel1927 Next



# Inter Annotator Agreement





## Conclusions

The ProMED-mail Coreference Corpus introduces a new domain of documents that can be used for training and evaluating existing coreference resolution methods and for developing new cross-domain coreference resolution methods.



## Acknowledgements

Resources and facilities at the  
VA Salt Lake City Health Care System



Funding support from the  
VA Informatics and Computing Infrastructure  
(VINCI), VA HSR HIR o8-204;

the Consortium for Healthcare Informatics  
Research (CHIR), VA HSR HIR o8-374;



Lawrence Livermore National Laboratory  
subcontract B573245; and

the Department of Homeland Security under ONR  
Grant N0014-07-1-0152.