



Information Extraction

From Latin American Terrorism Articles

Natural Language Processing

CS 6340 Fall 2017

Greeshma M

Sheetal Krishna MJ

System Design & Components

ID

- Used Regular Expressions to identify the news story identifier and to split the document into individual articles.

`((DEV-MUC3-)|(TST1-MUC3-)|(TST2-MUC4-))[0-9]{4}`

Incident

- Extracted the incidents from the answer keys and used the corresponding articles to train the classifier.
- Used Scikit Stochastic Gradient Descent (SGD) classifier.



Victim

- Used Stanford's NER parsers to identify the victim names that weren't terrorist organizations, etc.
- Obtained common victims by going through the answer keys.

Weapons

- Made a consolidated list of keywords using the identified weapons obtained from the answer keys of development and test datasets.
- Categorised them into weapons of different kinds and scanned the articles for the same.

Perpetrator Organization

- Similar to weapons extraction, we made a list of common perpetrator organizations that appeared in Latin American News, and grouped them together when they were referred to by common names. This helped avoid duplicate answers.
For example, [FMLN, Farabundo Martí National Liberation Front]



Emphasis / Originality

- Utilized simple heuristics to design our information system.
- The system is very specific to Latin American Terrorism News Articles.
- Ensured that there were no duplicate answers for weapons and organizations by categorizing them.
- We used semantic similarity for extracting weapons and incidents but didn't end up using it since the current techniques yielded better accuracy.



Performance

SCORES for ALL Templates

	RECALL	PRECISION	F-MEASURE
Incident	0.69 (69/100)	0.69 (69/100)	0.69
Weapons	0.70 (31/44)	0.15 (31/207)	0.25
Perp_Ind	0.00 (0/69)	0.00 (0/0)	0.00
Perp_Org	0.60 (28/47)	0.16 (28/178)	0.25
Targets	0.00 (0/62)	0.00 (0/0)	0.00
Victims	0.33 (47/144)	0.19 (47/245)	0.24
-----	-----	-----	----
TOTAL	0.38 (175/466)	0.24 (175/730)	0.29

MIDPOINT RESULTS

SCORES for ALL Templates

	RECALL	PRECISION	F-MEASURE
Incident	0.85 (85/100)	0.85 (85/100)	0.85
Weapons	0.67 (26/39)	0.46 (26/57)	0.54
Perp_Ind	0.00 (0/101)	0.00 (0/0)	0.00
Perp_Org	0.45 (19/42)	0.28 (19/68)	0.35
Targets	0.00 (0/59)	0.00 (0/0)	0.00
Victims	0.40 (59/149)	0.24 (59/249)	0.30
-----	-----	-----	----
TOTAL	0.39 (189/490)	0.40 (189/474)	0.39

FINAL RESULTS



Regrets and Successes

Regrets

- Our main regret was not being able to get a good accuracy on Perp-Indiv and Target after we tried using pattern recognition.

Successes

- The classifier used for Incident extraction performed better than expected (0.85 F-measure).
- Getting good F-measure for Prep-Organization.
- Finishing within the Top-5 :)



Team Member Contributions

Sheetal

Implemented the classifier for Incidents, worked on Weapons, Perpetrator Organisation and Perpetrator Individual.


Greeshma

Worked on figuring out the the method to classify Incidents, increased the accuracy by using a better classifier. Implemented Victim, Perpetrator Organisation and Targets.

We pair programmed most of the code.



External Resources

- [SciPy](#)
 - [Spacy](#)
 - [NLTK Toolkit](#)
 - [Sematch Tools](#)
 - [Wordnet Toolkit](#)
 - [Scikit-Learn Tools](#)
 - [Stanford NER Parser](#)
 - [Stanford POS Tagger](#)
 - [Stanford CoreNLP Tools](#)
 - [Information Extraction for MUC-3 Terrorism Domain Corpus](#)
- 

Thank you :)

