Deep versus broad methods for automatic extraction of intelligence information from text

Neil C. Rowe, Jonathan Wintrode, Jason Sparks, Jonathan Vorrath, and Matthew Lear

U.S. Naval Postgraduate School

ncrowe@nps.edu

Two strategies for intelligence extraction from text

- Deeper strategy
  - Use natural-language syntax to build model of the target information
  - Code it efficiently
  - Enter extracted data into a database
- Broader strategy
  - Use Web search engine to find relevant Web pages
  - Search for keywords on those pages to find relevant sentences
  - Use target phrase patterns involving proper nouns to rate the sentence candidates
  - Enter extracted data into a database

Example of deeper extraction: Location expressions

- This was part of the TEMPPTS Project at SPAWAR, with goal to give increased situational awareness about enemy locations.
- Intelligence reports give high-value low-volume information in natural language; can we automatically extract their location information for presentation in situational awareness tool?
- Approach focused on the grammar rules for location expressions. These were triggered by a set of location-indicating keywords (sometimes ambiguous, but multiple clues were used).

Example input to the location extractor

A FROG-7 BRIGADE WAS IN KUWAIT 40 KM WEST OF KUWAIT CITY AND ABOUT 5 KM NORTHWEST OF AL JAHRA.

WHILE THIS /BRIGADE IS MOST LIKELY ONE OF THE TWO IRAQI FROG-7 BRIGADES THAT DEPARTED THE AL MUFRASH AREA OF IRAQ, THE IRAQIS USE STAKEBED TRUCKS TO TRANSPORT FROG-7 AIRFRAMES.

THE FROG BRIGADE CONSISTED OF NINE TEL (AT LEAST THREE, AND POSSIBLY FOUR, LOADED WITH AIRFRAMES), FIVE STAKEBED TRUCKS (TWO WITH POSSIBLE ROCKETS), AND NUMEROUS SUPPORT EQUIPMENT.

AUGUST, AN ARTILLERY BATTALION WAS IN THIS AREA BUT DEPARTED.

TWO IRAQI FROG-7 BRIGADES WERE DEPLOYED IN AN OPEN AREA SOUTHEAST OF AL MUFRASH, IRAQ, AT 30-12N/047-33E.

THE SECOND IRAQI FROG BRIGADE LOCATED IN KUWAIT.

THE UNIT IS 5 KM NORTHEAST OF THE OTHER FROG BRIGADE.

MORE DETAILED INFORMATION TO FOLLOW.

Data found by our extractor for the example

Sentence #14 Subject [a frog-7 brigade] Verb [was] Link [in] Modifier [about] Object [kuwait] Time [] Coordinates [] Message Timestamp [1637z 03 july 1995] Weight 7.90748

Sentence #14 Subject [a frog-7 brigade] Verb [was in kuwait] Link [west of] Modifier [40 km about] Object [kuwait city] Time [] Coordinates [] Message Timestamp [1637z 03 july 1995] Weight 9.214765

Sentence #15 Subject [trucks] Verb [] Link [to] Modifier [most] Object [transport frog-7 airframes] Time [] Coordinates [] Message Timestamp [1637z 03 july 1995] Weight 911.0

Sentence #15 Subject [trucks] Verb [] Link [to] Modifier [most] Object [transport frog-7 airframes] Time [] Coordinates [] Message Timestamp [1637z 03 july 1995] Weight 911.0

Sentence #15 Subject [transport frog-7 airframes] Verb [] Link [] Modifier [most] Object [trucks] Time [] Coordinates [] Message Timestamp [1637z 03 july 1995] Weight 911.0

Sentence #16 Subject [five] Verb [stakebed trucks] Link [] Modifier [at possible] Object [] Time [] Coordinates [] Message Timestamp [1637z 03 july 1995] Weight 17.186014999999998

Data extracted, page 2

Sentence #17 Subject [an artillery battalion] Verb [was] Link [in] Modifier [but] Object [this area] Time [] Coordinates [] Message Timestamp [1637z 03 july 1995] Weight 7.90748

Sentence #18 Subject [iraqi frog-7 brigades] Verb [were deployed] Link [in] Modifier [] Object [an open area] Time [] Coordinates [at 30-12n/047-33e] Message Timestamp [1637z 03 july 1995] Weight 1.908545

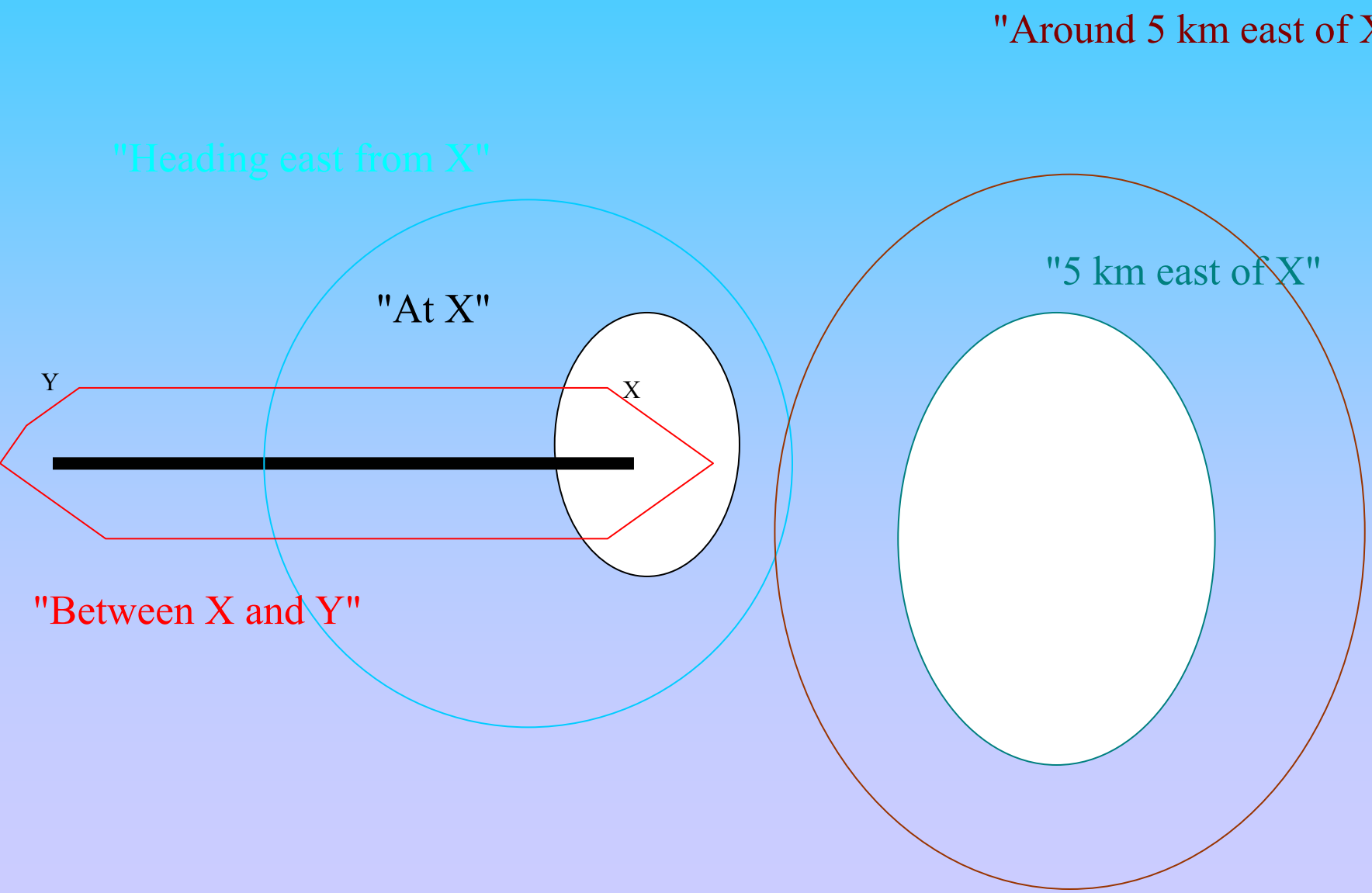
Sentence #18 Subject [an open area] Verb [] Link [southeast of] Modifier [] Object [al mufrash] Time [] Coordinates [at 30-12n/047-33e] Message Timestamp [1637z 03 july 1995] Weight 13665.0

Sentence #19 Subject [iraqi frog brigade] Verb [located] Link [in] Modifier [] Object [kuwait] Time [] Coordinates [] Message Timestamp [1637z 03 july 1995] Weight 7.90748

Sentence #19 Subject [iraqi frog] Verb [brigade located] Link [in] Modifier [] Object [kuwait] Time [] Coordinates [] Message Timestamp [1637z 03 july 1995] Weight 3.1520599999999

Sentence #20 Subject [the unit] Verb [is] Link [northeast of] Modifier [5 km] Object [the other frog brigade] Time [] Coordinates [] Message Timestamp [1637z 03 july 1995] Weight 229.31691999999998

Handling fuzzy location expressions



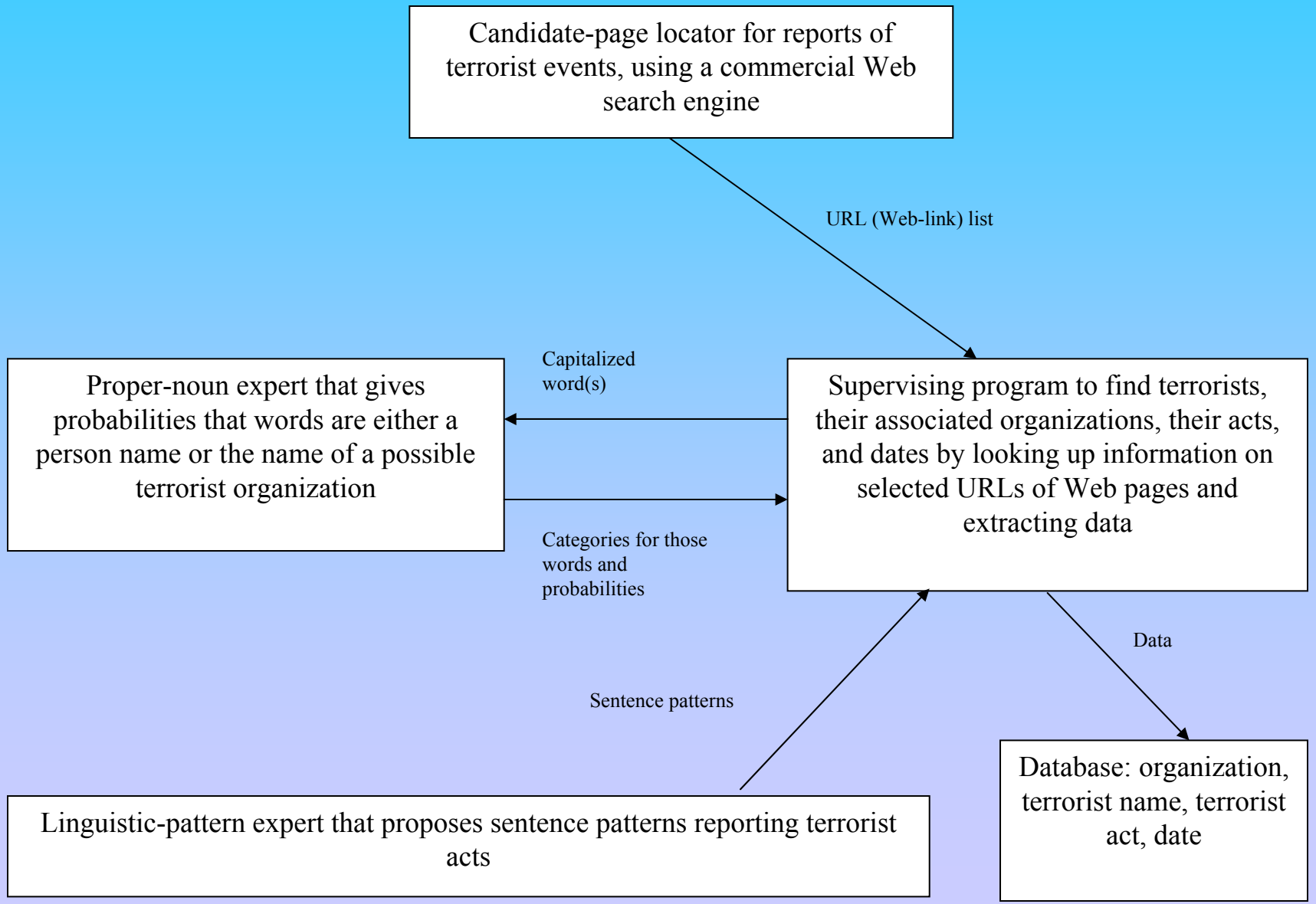
The broader approach: the terrorism-event extractor

Goal: Find sentences reporting terrorist acts and index their data. This would help show patterns.

Four parts:

- Search-engine (Alta Vista) lookup based on carefully crafted query of terrorism-related words
- Proper noun expert (for finding personal and organization names)
- Sentence-pattern expert (for finding sentence patterns suggesting terrorist acts)
- Overall control with Bayesian probabilistic rating of candidates and storing results in a table

Block diagram



The query to Alta Vista (candidate-page locator)

(news OR aggression OR attack OR assail OR assault OR barrage OR blast OR (blow AND up) OR bomb OR bombard OR bombardment OR bombs OR beheading OR burst OR (car AND bomb) OR (car AND bombing) OR cell OR (sleeper AND cell) OR crush OR damage OR decimate OR destroy OR detonate OR explosion OR explosive OR fire OR harm OR hostilities OR hurt OR IED OR (improvised AND explosive AND device) OR (islamic AND militants) OR infidel OR jihad OR kill OR kidnapping OR kidnap OR maim OR salvo OR shell OR strike OR (suicide AND bomber OR onslaught OR raid OR organization OR network OR terror OR terrorist OR terrorism OR briefing OR (homeland AND security) OR freedom OR violence OR guerrilla OR rebel OR rebellion) AND (NOT (editorial OR blog OR blogspot OR opinion OR puzzlers OR boggle OR puzzle OR wordlist))

Example keyword clues for rating sentences

Word	Number of Occurrences	Occurrences in a Terrorism-Related Sentence	Probability of Occurrence Given a Terrorism-Related Sentence	Probability of Occurrence
Qaeda	114	69	0.210	0.014
Al	277	99	0.302	0.034
Terror	314	96	0.293	0.039
Bomb	207	76	0.232	0.026
Zarqawi	16	14	0.043	0.002
Attack	180	42	0.128	0.022
Kill	186	34	0.104	0.023

Phrase	Number of Occurrences	Occurrences in a Terrorism-Related Sentence	Probability of Occurrence Given a Terrorism-Related Sentence	Probability of Occurrence
bin laden	37	10	0.030	0.005
the middle east	23	7	0.021	0.003
head of	25	6	0.018	0.003

Example terrorism-related phrase patterns

Angular brackets <> denote categories of words.

<organization> carried out <terrorist-act>
<organization> targeted <place>
<place> bombed by <organization>
<place> destroyed by <organization>
<person> assassinated <person>
<person> strike <place>
<organization> raid <place>
<person> hostage
<organization> ties to <organization>
bombing at <place>
<person> commands <organization>
<person> kidnapped
<person> freed

Clues for classifying proper nouns

- We need to identify person names (e.g. "John"), location names (e.g. "Baghdad"), and organization names (e.g. "Al Qaeda").
- We used lists from our previous natural-language research (MARIE-4) for all (2821, 433, and 269 words and phrases respectively).
- We also had a big list of known English words (about 30,000) that were not proper nouns, to rule out known words.
- We also experimented with automatic classification by using capitalized words following location prepositions like "at" and "in". We got 58% precision in identifying location proper nouns this way.

Example terrorism-related proper nouns

- **"Zarqawi** ordered bombings on targets near **Baghdad."**
  - Person/Org: **Zarqawi**
  - Location: **Baghdad**
- "A British Muslim has been captured in northern **Iraq** by **Kurdish** security forces after being suspected of fighting with the Islamic terror group **Ansar-al-Islam."**
  - Location: **Iraq**
  - Location: **Kurdish**
  - Person/Org/Location: **Ansar-al-Islam**

Example filtered sentence

- "A car ***bomb*** *exploded* in front of the Italian headquarters in the city of Nasariya, **kill**ing 19 Italians and 9 Iraqis, and jolting Washington and its allies away from what has been a deceptive illusion: the **guer**illa war is by no means limited to the notorious 'Sunni triangle' northwest of Baghdad."
- **Bold** indicates keywords matched by Naïve Bayes formula. *Italics* indicates phrase matched to a linguistic pattern, "bomb exploded".

Useful probabilities: p("bomb" | terror_related) = 0.232,
p("bomb" | ¬ terror_related) = 0.017,
p("kill"| terror_related) = 0.104,
p("kill" | ¬ terror_related) = 0.020,
p(terror_related) = 0.04, p(¬ terror_related) = 0.939.

Naïve Bayes Evaluation

$$p(terror\ related \mid sentence) =$$

$$1/(1 + g("kill", terrorrela\ ted)\, g("bomb", terrorrela\ ted)/o(terrorrela\ ted))$$

$$g(c,d) = p(c \mid \neg d)/\, p(c \mid d)$$

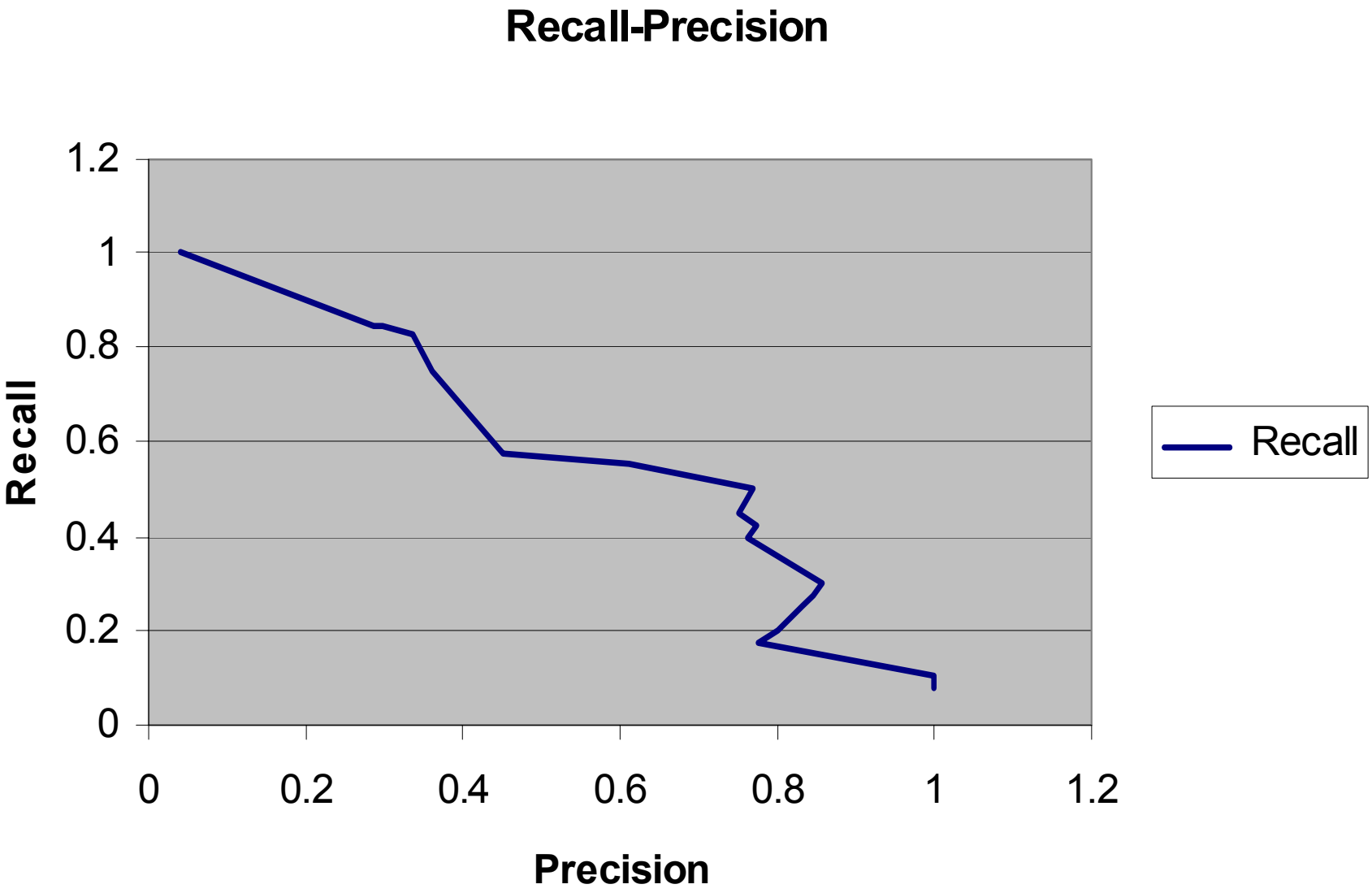
$$o(x) = p(x)/(1 - p(x))$$

$$\frac{1}{1 + \dfrac{0.020 \cdot 0.017 \cdot 0.959}{0.104 \cdot 0.232 \cdot 0.041}} = 0.754$$

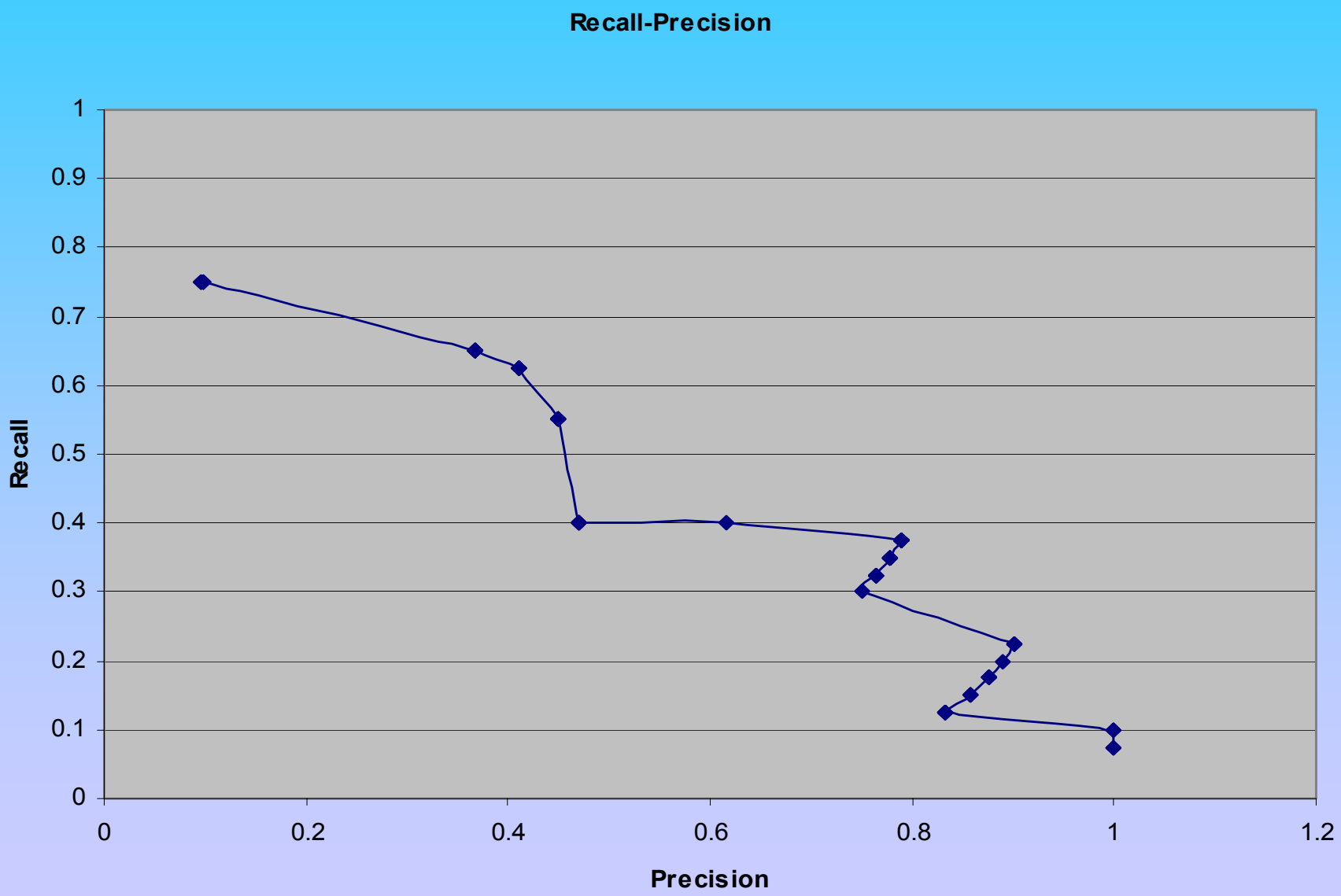
Training and testing

- A Naïve Bayes formula rated likelihood of a sentence being terrorism-related, from recognized words. Probabilities were obtained from training data.
- 260,000 (10%) of sentences were above our carefully-chosen threshold; others were ignored. (An improved system could assign probabilities to patterns and include in Naïve Bayes formula.)
- For more extensive tests, we started with 800 URLs and crawled to find 5500 related URLs.

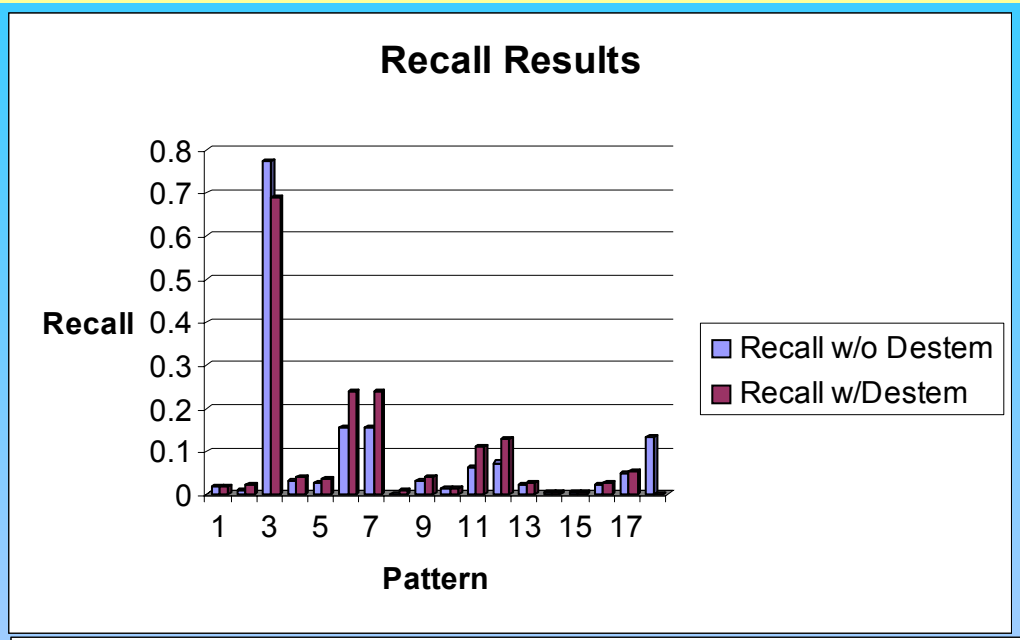
Overall recall-precision graph of TerrorPageCrawler

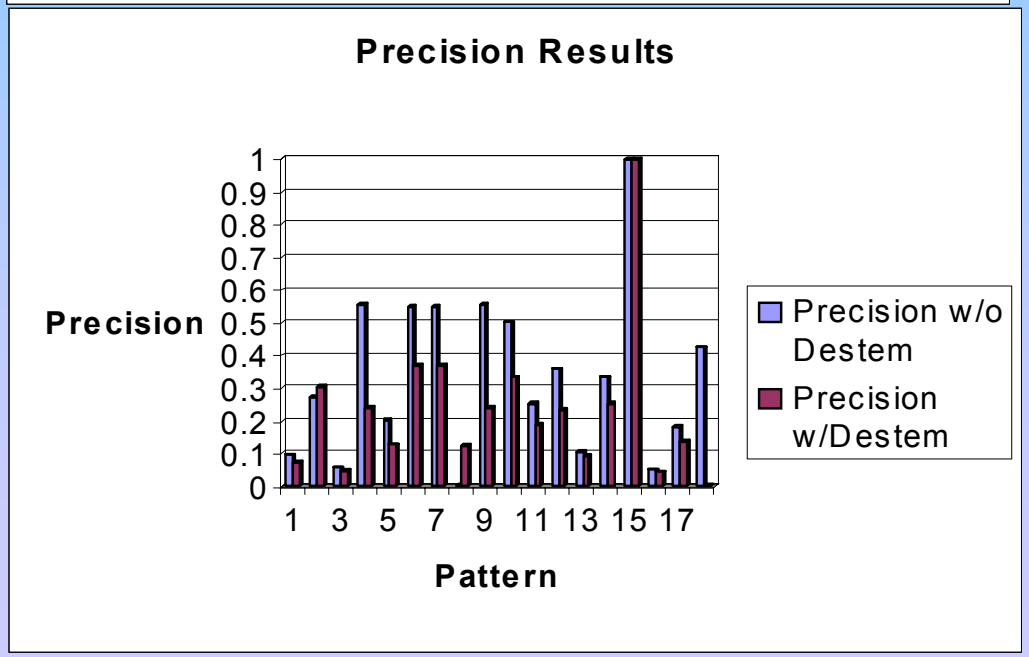


Poorer performance without sentence patterns



The value of destemming words before matching





Conclusions

- XML isn't necessary to do intelligence extraction.
- Google alone (or any search engine) is usually insufficient to do intelligence extraction from the Web -- additional filtering is necessary.
- The additional filtering can use grammar rules or sentence patterns -- and even simple ones can provide significant improvements.
- More work needs to be done on training taggers for parts of speech, a weakness in both projects.