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

Neil C. Rowe, Jonathan Wintrode, Jason Sparks, Jonathan Vorrath, and Matthew Lear
Code CS/Rp, 833 Dyer Road
U.S. Naval Postgraduate School
Monterey, CA 93943 USA
(831) 656-2462, fax (831) 656-2814
ncrowe@nps.edu, jwintrod@nps.edu, jsparks@nps.edu,
jjvorrat@nps.edu, mrlear@nps.edu

Abstract

Extraction of intelligence from text data is increasingly becoming automated as software and network technology increases in speed and scope. However, enormous amounts of text data are often available and one must carefully design a data mining strategy to obtain the relevant nuggets of gold from the mountains of useless dross. Two strategies can be tried. A "deep" approach is to use a few strong clues to find reasonable sentence candidates, then apply linguistic restrictions to find and extract key information (if any) surrounding the candidates. A "broad" approach is to focus on large numbers of weaker clues such as specific words whose implications can be combined to rate sentences and present those of high likelihood of relevance. In the work reported here, we tested the deep approach on military intelligence reports about enemy positions, which were relatively short text extracts, and we tested the broad approach on news stories from the World Wide Web involving terrorism, which presented a large volume of text information.

1. Extracting location information from intelligence reports

Intelligence reports of many kinds are routinely sent to operational military organizations. While some of this data can be and is formatted into short labeled fields, the unpredictable nature of military challenges necessitates a significant amount of natural-language description to provide fuller explanations of formatted data and unanticipated phenomena of interest. Natural language is also easier for untrained personnel to use than formatted data, and its use can reduce errors in encoding and decipherment of formatted data.

In particular, a significant number of intelligence messages arrive on military ships every day. A hope is that this information could be subject to data mining to automatically extract and format a critical kind of information, locations of military targets. Such a facility, if sufficiently fast, could provide valuable input to mission planning. It could define probability distributions of targets within which aerial analysts or pilots could try to find them and strike them. We explored these issues as part of the TEMMPTS project at Navy SPAWAR, San Diego, California.

1.1. Extracting location expressions

Natural language descriptions permit qualifications and conditions to be attached to target descriptions like "probably at", "recently at", "was seen at", "could be at", "tend to be at", all of which have different implications for targeting procedures. Our goal was to use modern ideas of

data mining to provide focused analysis of the text without doing complete natural-language processing of all the sentences, since processing speed is critical with targeting intelligence. (Morimoto et al, 2003) had similar goals for mining of location expressions in a more general class of text. We used the language Java to implement a prototype program since it is efficient and widely available.

Our approach is to do "partial parsing" of the text, parsing centered around just the words that are strong indicators of location expressions. This permits us to ignore large blocks of irrelevant text. For location expressions, the strongest clues are location prepositions ("at", "beside", "in", etc.) and location verbs ("located", "heading", "deployed", etc.). We thus scan text first to find such words. We also seek obvious location formats like latitude-longitude expressions; dates and times are also important concomitant information for locations, so we look separately for their formats too. Special formats almost invariably apply globally to the entire sentence in which they occur, as we rarely saw instances of more than one coordinate pair or time in a sentence.

Recognizing the words requires a comprehensive dictionary with taxonomic information since our input is unrestricted text. We used the dictionary from the MARIE-2 full parser (Rowe, 2004) that was designed for another military application, the captions at the Naval Air Warfare Center in China Lake, California. Much of this dictionary came from the Wordnet thesaurus system (Miller et at, 1990) plus 991 code formats, 2818 person names, 423 place names beyond those in Wordnet, 1199 common misspellings, 879 common abbreviations, and 3234 words we explicitly defined to cover the remaining words in the full set of 30,000 unclassified captions provided by China Lake. This dictionary was preprocessed and stored in hash tables for quick retrieval. We also obtained, from MARIE-2 statistics on the captions, frequencies of occurrence of word senses and grammar rules, which we then used to help rate possible phrase extractions in cases of ambiguity.

Once we recognize clue words, we try to extract subject and object noun phrases. To avoid real-time parsing, we created hash tables of all possible noun phrases, verb phrases, and participial phrases of one to four words consistent with MARIE-2's grammar of 263 two-replacement and one-replacement rules (i.e. in Chomsky normal form). We scan for all three before location prepositions and around location verbs; we scan for noun phrases after both prepositions and verbs. Then, to get around the limitation of a maximum of four words in the hash table, we iteratively examine words before preceding noun phrases and after following noun phrases to extend them. We append only nouns in extending to the right; we append only nouns, adjectives, and articles (in that order) when extending to the left.

However, not just any grammatical phrase containing a location preposition or verb is accepted for indexing. The principal verb of a verb phrase must be a location verb. The principal noun of a subject noun phrase must have at least one sense that represents something of military interest: a military unit, a vehicle, a building, or a military activity. The principal noun of an object noun phrase must be either something of military interest or a location. We use the Wordnet hierarchy to classify nouns. These classifications are applied to the headword of the phrase as determined by the MARIE-2 parsing rules. Subject-verb headword agreement in number is required.

For example, consider "A nuclear weapons laboratory of Hussein's was built near Rawanduz". "Near" is a location preposition and "built" is a location verb. "Was built" is the only possible choice of a verb phrase before the preposition; "a nuclear weapons laboratory of Hussein's" is the subject noun phrase, and "Rawanduz" is the object noun phrase of the preposition. We would recognize "weapons laboratory of Hussein's" as a four-word noun phrase that could be extended leftward with adjective "nuclear" and determiner "a". "Laboratory" is the headword of the subject

phrase, and it is a building in one sense, and thus a possible military target; "Rawanduz" is a geographical place name, and thus a possible object location.

1.2. Example

Here is an example input text, from an unclassified test example provided by Navy SPAWAR in imitation of real messages. Note only the "RMK" field will be subject to natural-language processing, but a few things can be captured from the formatted information too.

ISSN 8741/U39MSG
R 031637Z JUL 95
EZ03
U N C L A S EFTO----------SECTION 01 OF 01 SECTIONS
EZ04
EZ05
MSGID/IIR/PMOE//
ITEMTYP/SIGNIFICANT ACTIVITY OR OB ITEMS//
ITEM/010/SILKWORM TEST FACILITY
DEP/IBE:00001ZZ0001/CAT:00001
/CTY:VX/Y//
LOC/GEO:335000N0811500W/UTM:01ZZZ00010001//
OTID/LTH:00001M/WTH:00001M/AZM:001/ELE:00001M/CMX:01L00001LL//
STATACT/OCC//
RMK/SUBJECT: PERSIAN GULF CRISIS UPDATE
/
/NOT FINALLY EVALUATED INTELLIGENCE
/
/TO FACILITATE ELECTRONIC ACCESS, THIS DOCUMENT HAS BEEN
/REFORMATTED TO ELIMINATE INFORMATION THAT DOES NOT PERTAIN
/TO GULF WAR ILLNESS ISSUES OR THAT IS CLASSIFIED. A COPY OF
/THIS REDACTED DOCUMENT, IN ORIGINAL FORMAT, IS AVAILABLE ON
/REQUEST.
/
/AUG 90
/
/TABLE OF CONTENTS
/1. IRAQ-KUWAIT BORDER DISPUTE
/ PERSIAN GULF CRISIS UPDATE
/ --PROBABLE IRAQI FROG IN KUWAIT
/COMBINE; COMPLETE
/1. IRAQ-KUWAIT BORDER DISPUTE
/PERSIAN GULF CRISIS UPDATE
/ -- PROBABLE IRAQI FROG IN KUWAIT
/
/A PROBABLE IRAQI FROG BRIGADE WAS IN KUWAIT
/
/IN OTHER ACTIVITY, IRAQ HAS REPOSITIONED SOME OF ITS
/FORCES IN SOUTHERN KUWAIT, NEAR THE SAUDI ARABIAN-BORDER. IN
/KUWAIT CITY, A CONVOY OF UNIDENTIFIED VEHICLES
/NEAR THE US EMBASSY

/THE FOLLOWING SIGNIFICANT ACTIVITY WAS IN
/ KUWAIT
/(SIWN)
/
/PROBABLE IRAQI FROG BRIGADE
/
/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.
/
/IRAQI GROUND FORCES DEPLOYED IN THE SOUTH
/
/A BM-21 MULTIPLE ROCKET LAUNCHER BATTERY HAS DEPLOYED ALONG
/THE COAST ROAD, 7 KM NORTH OF THE SAUDI ARABIAN-KUWAITI
/BORDER.
/
/IRAQI AIR FORCES
/
/ - AN IRAQI CANDID HEAVY JET TRANSPORT HAS ARRIVED AT ALI AL
/ SALEM AIRFIELD, IN CENTRAL KUWAIT
/
/ A PROBABLE
/ KUWAITI MIRAGE FL MULTIROLE FIGHTER WAS PARKED IN FRONT OF A
/ HARDENED AIRCRAFT BUNKER AT THE AIRFIELD.
/
/END OF MESSAGE
/1.5 (c)
/93692-93692\fR

Here is the output of our program from processing this text. Note it produces a series of fields which may or may not be filled depending on what is available in the text. "Timestamp" comes from the header of the message.

Sentence #8 Subject [probable iraqi frog] Verb [] Link [in] Modifier [] Object [kuwait] Time [] Coordinates [] Message Timestamp [1637z 03 july 1995] Weight 455.5
Sentence #9 Subject [probable iraqi frog] Verb [] Link [in] Modifier [] Object [kuwait] Time [] Coordinates [] Message Timestamp [1637z 03 july 1995] Weight 455.5

Sentence #10 Subject [a probable iraqi frog brigade] Verb [was] Link [in] Modifier [] Object [kuwait] Time [] Coordinates [] Message Timestamp [1637z 03 july 1995] Weight 7.90748
Sentence #11 Subject [forces] Verb [] Link [in] Modifier [some near] Object [southern kuwait] Time [] Coordinates [] Message Timestamp [1637z 03 july 1995] Weight 6832.5
Sentence #12 Subject [vehicles] Verb [] Link [near] Modifier [] Object [the us embassy] Time [] Coordinates [] Message Timestamp [1637z 03 july 1995] Weight 216362.5
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
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
Sentence #23 Subject [forces] Verb [deployed] Link [in] Modifier [] Object [the south] Time [] Coordinates [] Message Timestamp [1637z 03 july 1995] Weight 118.6122
Sentence #23 Subject [iraqi ground] Verb [forces deployed] Link [in] Modifier [] Object [the south] Time [] Coordinates [] Message Timestamp [1637z 03 july 1995] Weight 0.035292
Sentence #24 Subject [launcher battery] Verb [has deployed] Link [along] Modifier [] Object [the coast road] Time [] Coordinates [] Message Timestamp [1637z 03 july 1995] Weight 1.908545
Sentence #27 Subject [transport] Verb [has arrived at] Link [] Modifier [] Object [ali al salem airfield] Time [] Coordinates [] Message Timestamp [1637z 03 july 1995] Weight 906.558875
Sentence #28 Subject [a probable kuwaiti mirage fl multirole fighter] Verb [was parked] Link [in] Modifier [] Object [front] Time [] Coordinates [] Message Timestamp [1637z 03 july 1995] Weight 7.63418
Sentence #28 Subject [aircraft bunker] Verb [] Link [at] Modifier [] Object [the airfield] Time [] Coordinates [] Message Timestamp [1637z 03 july 1995] Weight 455.5
Sentence #28 Subject [aircraft] Verb [bunker] Link [at] Modifier [] Object [the airfield] Time [] Coordinates [] Message Timestamp [1637z 03 july 1995] Weight 3178.80696

1.3. From location descriptions to probability distributions

Once we have extracted location expressions, we need to interpret their semantics to use them in mission planning. A key problem with many targets such as military units and vehicles is that they are mobile, and their locations may change from the time they are observed (Custy, McDonnell, & Gizzi, 2002). Iraqi Scud mobile missile launchers during the Gulf War were an example: The flash of their launch was easy to see by satellite, but immediately afterward the launcher would depart from the site in some unpredictable direction.

In general we can postulate a probability distribution around observed locations of mobile targets for their current location. For most purposes, this can be radially symmetric about the original location, and represents the distance that the target could have moved. It will increase in size roughly proportionate to the amount of time between observation and mission. The constant of proportionality depends on the speed of the target: Vehicles can travel at road speeds, while military personnel on foot can travel at foot speed.

The probability distribution must also reflect the fuzziness of the natural-language location description. "Probably", "approximately", and "near" when used as adverbs imply probability distributions for the original distribution that must then be convolved with a movement distribution to obtain the full distribution. "Along the coast road" has a different kind of fuzziness since a road occupies a wide range of space. An expression like "5 km northeast of the other Frog brigade" has three kinds of associated fuzziness since the brigade occupies some area, and even if you take its center of gravity as its precise location, "5 km northeast" allows some degree of uncertainty in latitude and longitude as well as in distance. All these error distributions need to be convolved to get a cumulative distribution. Figure 1 shows some examples. Note that these may also apply to nonmobile targets when intelligence about them is not certain.

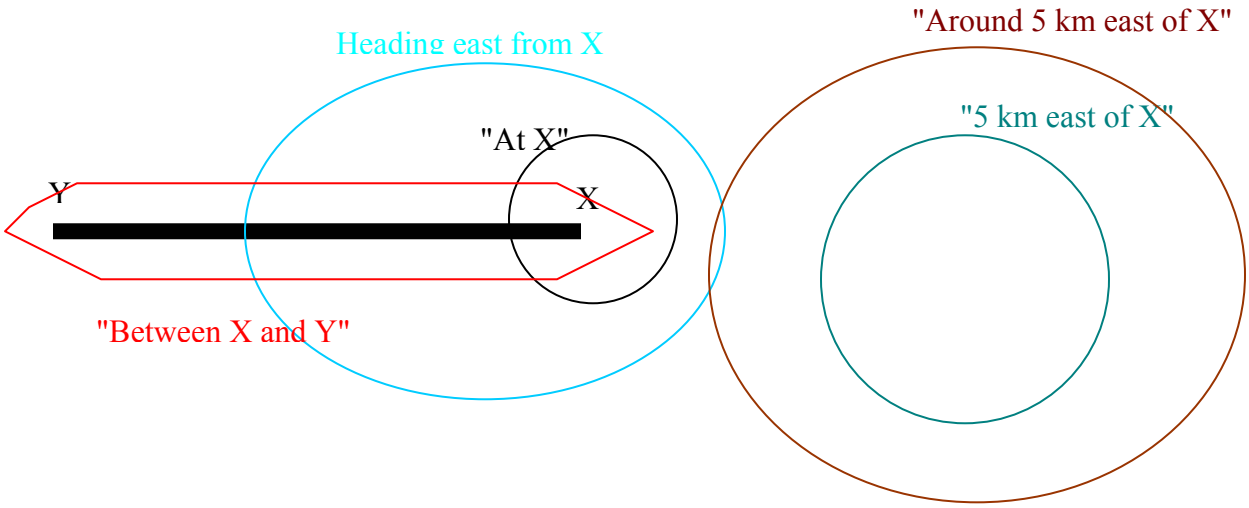

Figure 1: Example linguistic spatial expressions.

Most distributions for fuzzy phenomena can be modeled as initially radially symmetric. However, an interesting class of expressions imply asymmetric distributions, e.g. "heading west". For these we need specially formulated distribution shapes with associated formula for expansion over time.

Radially symmetric distributions for movement make sense in isotropic terrain where travel is equally easy in any direction. The ideal case would be a military unit on foot in an unobstructed plain. But road networks provide much faster movement along narrow corridors, and a more sophisticated modeling could take this into account to provide irregular anisotropic distributions (Rowe, 1997). A good model for many cities is a rectangular grid of roads. Then a city-block (Hamming) metric is necessary to measure distance. On a square grid, the city-block metric would give an effort proportionate to |cos(t)|+|sin(t)| to travel at angle t to the grid, so the probability distribution would have width the inverse of that at bearing t from its center.

Still more sophisticated distributions can be created in the case of areas with few roads. Here the speed of travel is much faster along roads than across other terrain. If the observation point of a vehicle is on such a road, its associated probability distribution will be diamond-shape at future times, since the best way to get anywhere is to follow the road for some distance and then cut away from it at a particular "critical angle" analogous to the angle of total internal reflection in optics (Rowe, 1997). Where the road bends the diamond pattern bends too, and at road intersections the diamond is replicated in every possible direction. See Figure 2, where the area inside the purple lines is the reachable area from the start in a fixed amount of time.

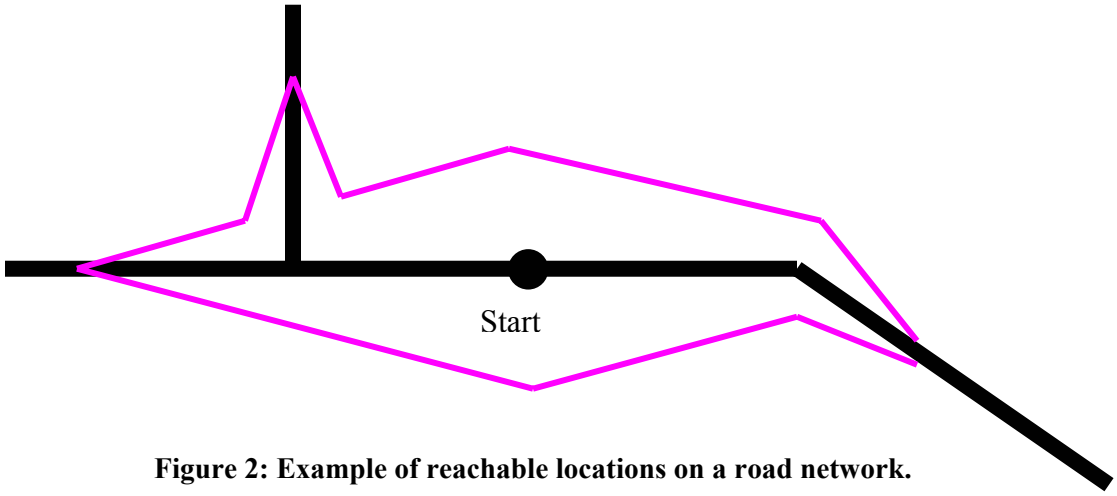

Figure 2: Example of reachable locations on a road network.

2. Extracting reports of terrorist acts from the World Wide Web

To explore a broader approach to extraction of intelligence information from text, we built a prototype system for automatically mining the World Wide Web for news reports of terrorist actions. For instance, we would like to recognize as relevant a news-report sentence like "The freight train was carrying military supplies near Fallujah west of Baghdad when an improvised bomb set four containers ablaze" and recognize as irrelevant "Sequel bombs at box office". The goal was to automatically build a database from which intelligence analysts could study to find trends. Our program is written in Java. The overall design of our system is shown in Figure 3.

Obtaining information about terrorist activities is an important task for intelligence, so automated assistance could be valuable (Popp et al, 2004). Many terrorist organizations have a flexible

distributed structure, so their internal communications are difficult to identify. In addition, manual collecting and analyzing the data about terrorists, their organizations, and their actions is quite tedious. It would be worthwhile to automate the tracking of terrorist networks by exploiting the clues inadvertently available in the vast amount of news information available on the Web, as a form of Web content mining (Kolari and Joshi, 2004). This idea is being explored currently in research sponsored by the U.S. Department of Homeland Security. (Gruenwald, McNutt, & Mercier, 2003) provides a similar approach to ours to the problem but with differences in emphasis and design.

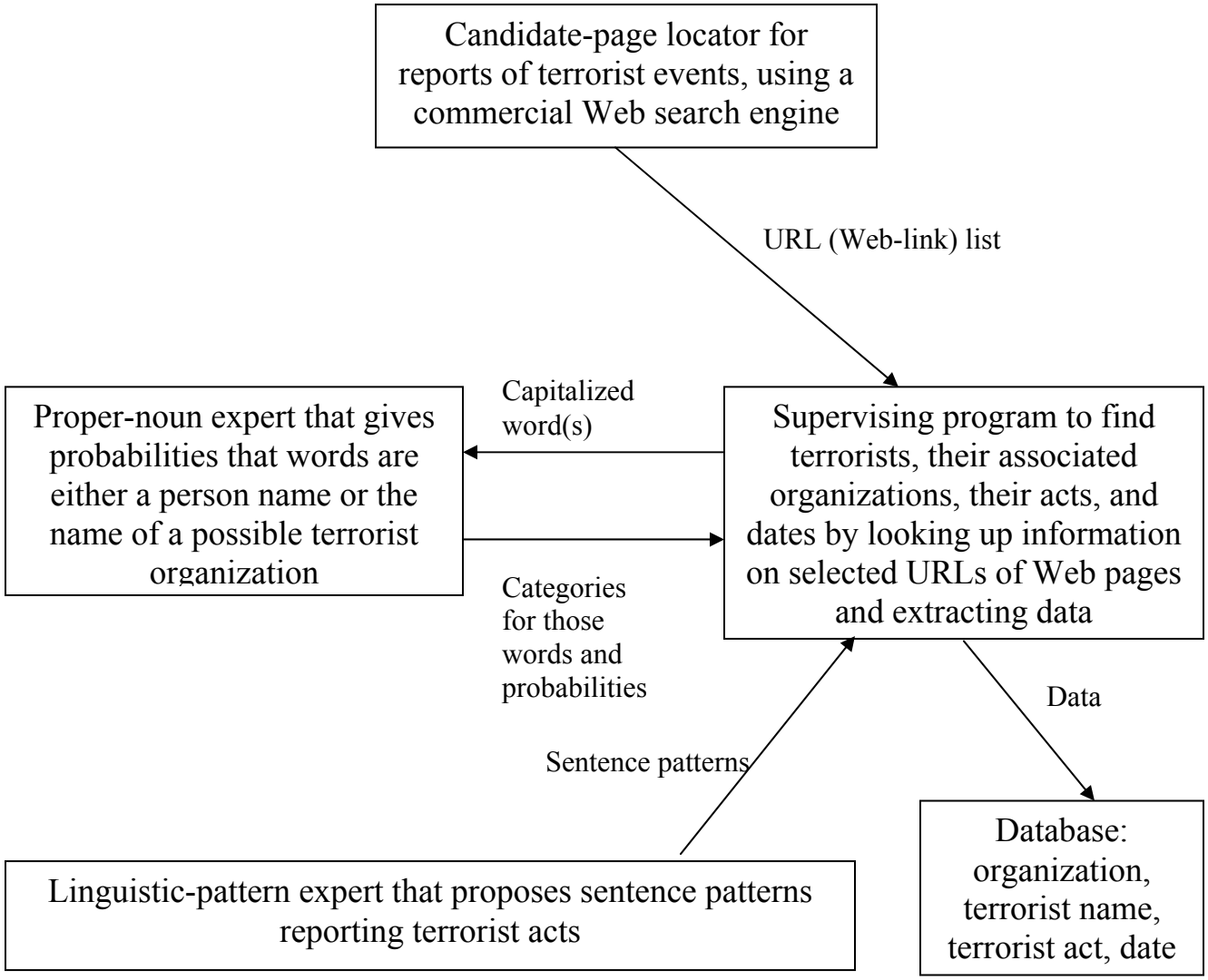

Figure 3: Organization of the terrorism-event Web page crawler.

2.1. The candidate-URL locator

A first step is to locate potentially useful Web pages. We used the Alta Vista search engine (www.altavista.com). Browsers like Alta Vista match a set of keywords to pages and sort output by a decreasing order of rating (McCurley and Tomkins, 2004). The rating uses a "secret formula" that incorporates factors such as the number of keyword occurrences on a page (Baerza-Yates, 1998), where the matches occur, and the number of links to the page.

We carefully constructed a query of terms related to terrorist attacks. We chose generic search terms rather than specific ones like "Laden" and "Qaeda" to provide robustness of our methods over time and for new threats. After some experiments, we used the query:

> (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))

Most of the query terms came from intuition, but important ones came only from studying results of simpler queries. Search terms were disjunctively connected since most were sufficient clues by themselves, with the exception of a few words like "car" and "bombing" that only are clues when used together. The negations in the query were added to prevent false positives for editorial, "blog", puzzle, and word-list sites. Note that we focused narrowly on collecting information about terrorist acts themselves rather than analysis of them. False positives appeared to be less a problem for the Google browser, in some manual experiments; however, at the time of these experiments, Google could not be used for free by a program.

A Java program sent the query to Alta Vista and then processed the top 1000 links that Alta Vista returned. Some subsequent elimination of pages was done (134 of the 1000) since Alta Vista was not reliable in eliminating pages including negated terms. For an initial assessment, fifty retrieved links we chosen randomly and manually inspected for information about terrorist attacks. 40% contained information specifying terrorist attacks, and as expected, the "com" and "org" domains were more fruitful than the "edu" and "mil" domains. For a more comprehensive assessment, the later stages of our system described below passed back data on 5500 pages they found starting with the 850 pages Alta Vista found, by looking for sentence structures that matched the forms of statements about terrorist acts. They also found very nearly a 40% success rate, using an intuitive notion of "success" (not the same as that in section 2.4).

2.2. The proper-noun expert system

Our proper-noun expert module recognizes and classifies proper nouns. Two related projects of (Barcala et al, 2002) and (Petasis et al, 2000) have addressed the problem of recognizing proper nouns; the latter focused particularly on the valuable idea of learning them automatically for specialized domains. Names of terrorist organizations and other terrorism-associated words were found by Web searches, and personal names came from the list in (Rowe, 2004) with some Arab-name additions.

Processing took as input a string of words selected from a Web page as will be explained in section 2.4; this was a sentence or part of a sentence that had some terrorism-related word. First, it compared each word of the string against a list of person-name words like "John", and it compared sequences of capitalized words against a list of terrorist-organization names. Comparison ignored case since capitalization is not uniform on Web pages. If the word is in either list, the word, the name of the list, and a probability of 1.0 are sent to the output. Otherwise, if K nearby words were terrorism-related, a probability of $K^2/(2 + K^2)$ was assigned to each capitalized word as being either a useful name or organization.

A recall-precision graph for proper-noun classification is shown in Figure 4, created by varying the threshold probability for a sample of test sentences we built ourselves. (Recall is the fraction of proper nouns correctly categorized as terrorism-related, of all those in the sample that were terrorism-related; precision is the fraction of proper nouns correctly categorized as terrorism-related, of all those identified as terrorism-related.) False alarms resulted when terrorist-related words had multiple meanings, like "hostile" which can refer to either a work environment or a corporate takeover. In most cases only one terrorism-related context word was associated with a proper noun, so counting their number is not helpful.

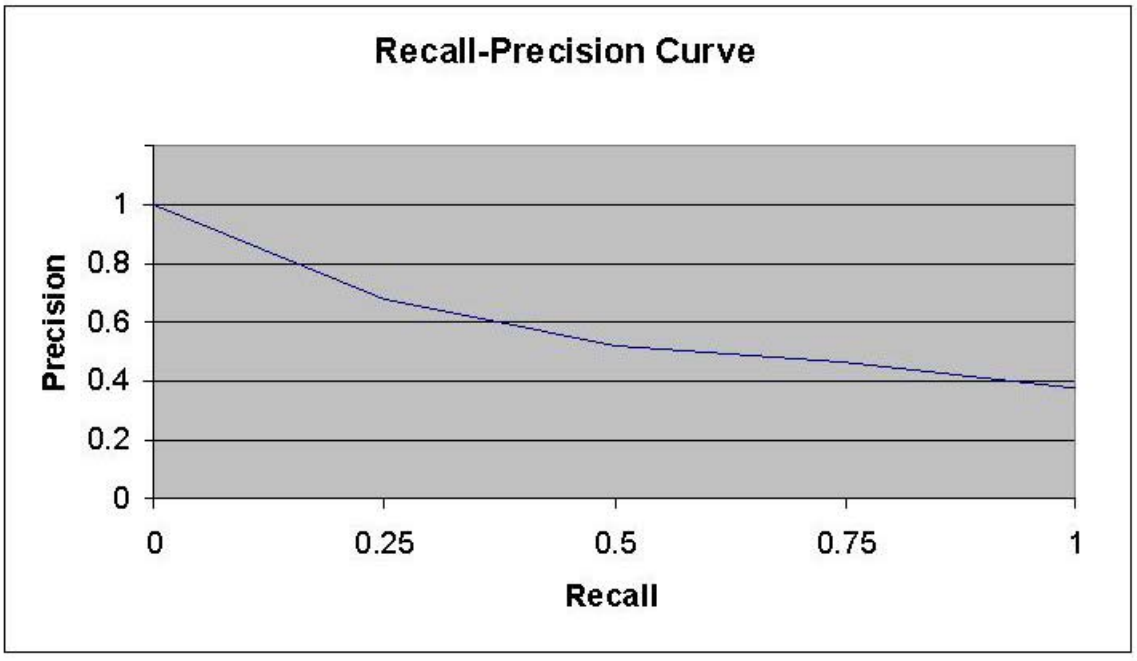

Figure 4: Recall-precision curve for effect of context on proper-noun identification.

Destemming and syntax information can also help categorize proper nouns. We had 3,163 unknown words of 44,376 in our test dataset, which was reduced to 1,221 possible names of people and organizations from reasoning about suffixes stripped by the destemmer. We also were able to make rough distinctions between names, organizations, and locations based on preceding prepositions such as "at", "in", and "nearby." For the 1500 sentences identified by the TerrorPageCrawler, this simple idea identified 1,022 words as locations, of which 586 were in fact locations, for a precision of 58%. These ideas could be used to improve the accuracy of the full system, but we did not have time to test them thoroughly.

2.3. The linguistic-pattern module

The linguistic-pattern module proposes sentence patterns to the main program which are suggestive of reports of terrorist events. Similar work is PALKA (Kim and Moldovan, 1995) which automatically creates patterns from training sentences that can then be used to extract information from additional sentences. Work on automatic tagging of grammatical categories in sentences (Brill, 1998) is also related since tagging represents a deeper sentence structure.

We used Backus-Naur Form to describe patterns, in which angular brackets indicate categories; Figure 5 lists our patterns. An initial subset were created by intuition, then additional patterns were created from test runs for terrorism-related sentences unmatched by the first set. Matching of patterns was done by first matching the explicit words, then the angular-bracket categories. A destemmer from MARIE-4 (Rowe, 2002) removed suffixes on morphologically related words, and performance improved though there was considerable variation in effectiveness (see Figures 6 and 7). Here recall means the fraction of relevant documents retrieved of all those in the pages found, and precision means the fraction of relevant documents retrieved of all those retrieved. The abovementioned proper-noun checker handled person and organization names, and terrorist acts were defined by a list of words.

<person> ordered <organization>
<organization> carried out <terrorist-act>
<organization> targeted <place>
<organization> murder <person>
<place> bombed by <organization>
<place> targeted by <organization>
<person> killed <person>
<person> shot <person>
<person> strike <place>
<organization> raid <place>
<person> plot
<person> charged
<person> arrested
<person> accused
<person> captive
<person> member of <organization>
<person> spokesman for <organization>
<person> head of <organization>
<organization> ties to <organization>
<person> released
<person> released by <organization>
<person> successor to <person>
<person> commands <organization>
<person> control <organization>
<person> kidnapped
<person> hijacked
<person> support <organization>
<organization> recruit
<person> exploded at <place>
<person> instructed <person>
<terrorist-act> in <location> in <time>
<organization> targeted <person>
<organization> bombed <place>
<place> destroyed by <organization>
<place> blown <organization>
<person> assassinated <person>
<person> attack <place>
<place> hit by <organization>
<person> plan
<person> surrenders
<person> captured
<person> indicted
<person> threaten <person>
<person> hostage
<person> spokesman for <organization>
<person> leader of <organization>
<organization> led by <person>
bombing at <place>
<person> released by <person>
<person> assumes leadership of <organization>
<person> president of <organization>
<organization> founded by <person>
<person> beheaded
<person> freed
<person> supply <organization>
<organization> training
<person> detonated at <place>

Figure 5: Sentence patterns indicating terrorism-related sentences.

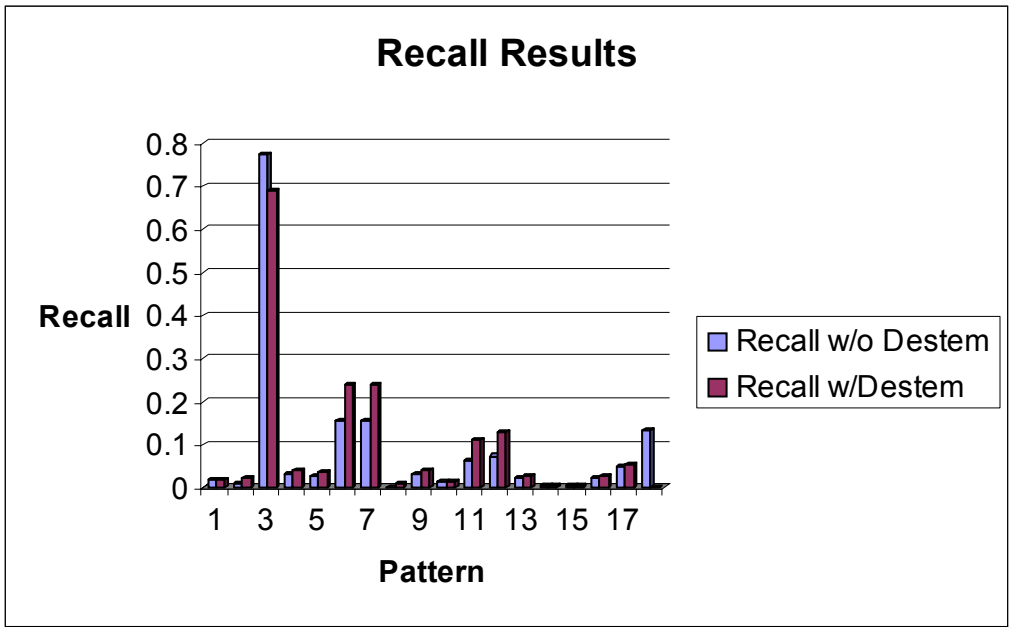

Figure 6 :The recall results before and after using the Destemmer.

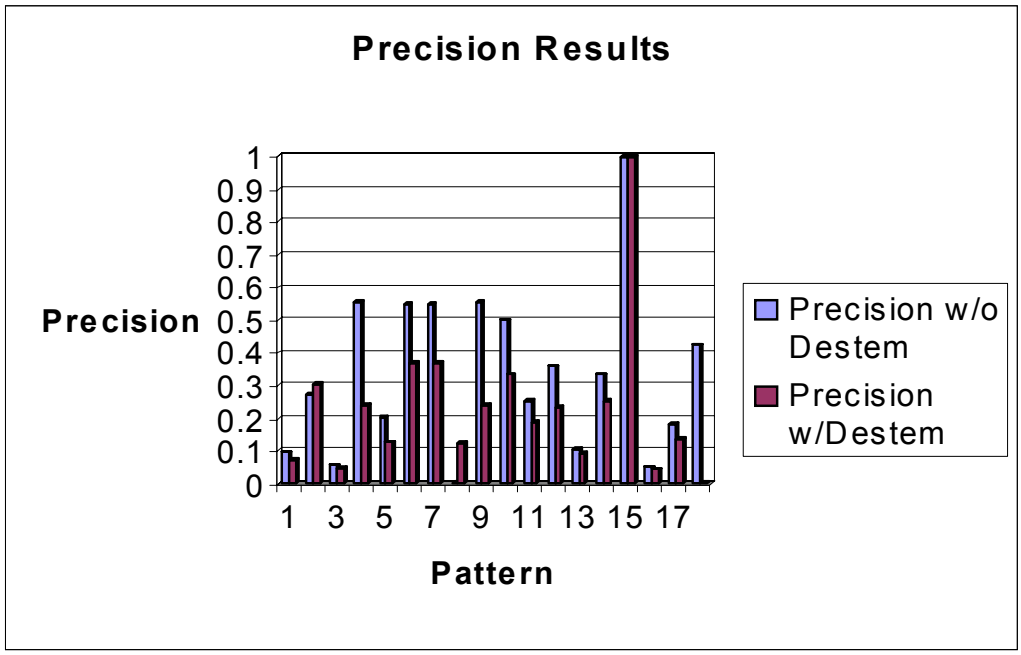

Figure 7: The precision results before and after using the Destemmer.

For each match of a pattern to a sentence, a "terror probability" was computed by awarding 0.5 for matching the explicit words and 0.25 for matching the words before and after explicit words. If the terror probability equalled or exceeded the threshold, the sentence was considered a match. Putting the threshold at 0.75 gave 46 terror sentences (14.7%), while setting it to 0.50 gave 283 terror sentences (90.7%). Figures 8 and 9 show representative recall-precision graphs of two

patterns. The overall recall and precision of the linguistic-pattern module were tested with the full system assembled as we will discuss in the next section.

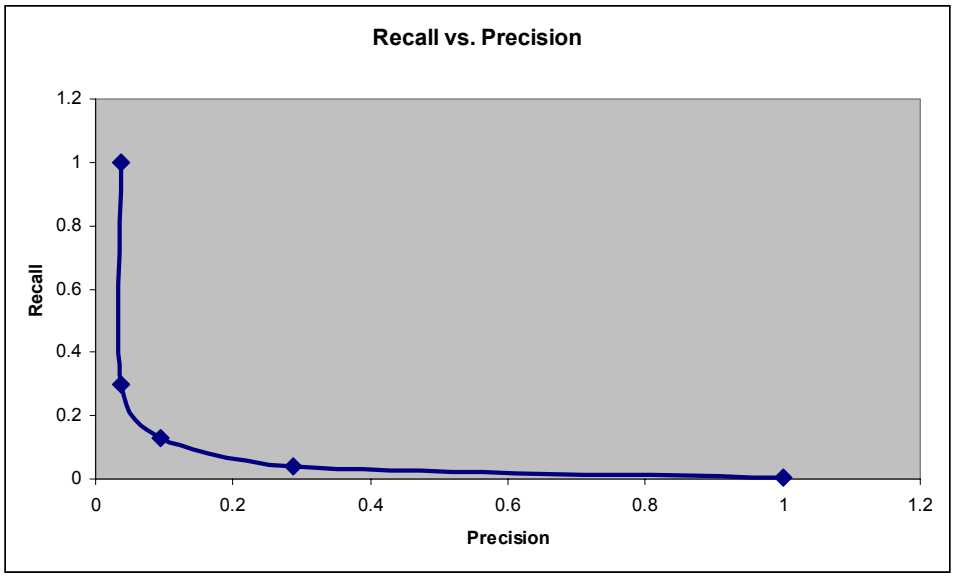

Figure 8: Recall-precision graph for "<person> killed <person>".

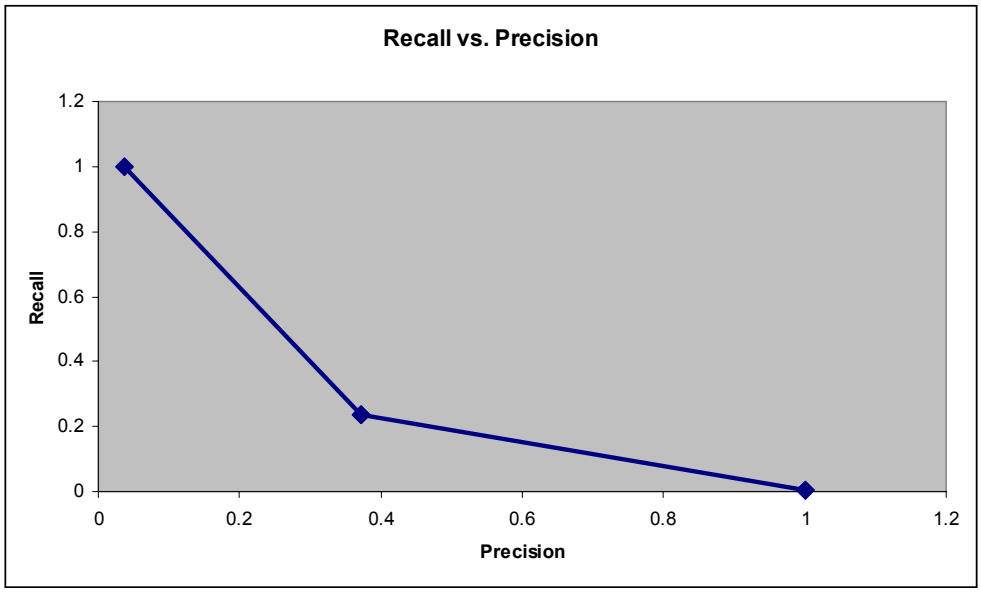

Figure 9: Recall-precision graph for "Bombing at <place>".

2.4. The supervisor program

The three modules described above are used by a supervisory module, the TerrorPageCrawler. It examines pages found by the candidate-URL locator, and extracts sentences that match patterns of the linguistic-pattern module, with the proper-noun expert identifying the key proper nouns. Besides what has been mentioned, it has:

- A "crawler" (or "spider") for systematically inspecting Web sites;
- A utility to extract sentences from HTML content;
- A machine-learning process to identify words indicating terrorist acts;
- A utility that filters out sentences based on their words;
- Linguistic-pattern matching including checking of proper nouns; and
- A Web interface to the results.

The crawler extends one from (Rowe, 2002) using threading code from (Heaton, 2002). Each worker thread obtains a URL from the queue, visits the URL, processes its content, and adds any new URLs it discovers to the queue. Sentence processing extracts the text from HTML input by removing the formatting tags.

Examination of a sample of 8077 sentences on pages returned by the crawler using the candidate-URL locator revealed that only 328 contained terrorism-related information. Statistics were collected on one, two, and three-word phrases in the sample. We determined words and phrases that occurred significantly more often in terrorism-related sentences than non-terrorism sentences, to use as clues. Tables 1 and 2 show the strongest indicators.

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

Table 1. Strongest one-word clues.

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

Table 2. Strongest phrase clues.

To obtain the probability that a sentence not previously seen is terrorism-related we used a Naïve-Bayes formula for calculating the conditional probability given the words and phrases in the sentence, assuming a-priori independence of the clue words:

$$P(TerrorSentence \mid clue_1, clue_2 ... clue_n) = \frac{P(clue_1 \mid TS) \cdot P(clue_2 \mid TS) \cdot ... \cdot P(clue_n \mid TS) \cdot P(TS)}{P(clue_1) \cdot P(clue_2) \cdot ... \cdot P(clue_n)}$$

To test the Bayesian learning process we ran the crawler to obtain an additional 1000 sentences, 41 of which were terrorism-related. We calculated conditional probabilities for those sentences using the formula. The results revealed that 88% of the sentences had a less than 0.1 likelihood of being terrorism-related. A recall-precision analysis is shown in Figure 10. It shows that accepting sentences above the 0.1 threshold results in a recall of 0.85 for the remaining 10% of candidate sentences. Here recall means the ratio of the number of terrorism-related sentences found to the number of terrorism-related sentences in the sample; precision means the ratio of the number of terrorism-related sentences found to the number of sentences found.

We incorporated these word probabilities into the crawler to rate candidate sentences and discard those below a threshold. Based on the results of the recall-precision analysis, a threshold of 0.10 was chosen, at which 90% of sentences were discarded while still maintaining a high recall. All data that the search finds is recorded in a flat-file database as a resource for further analysis.

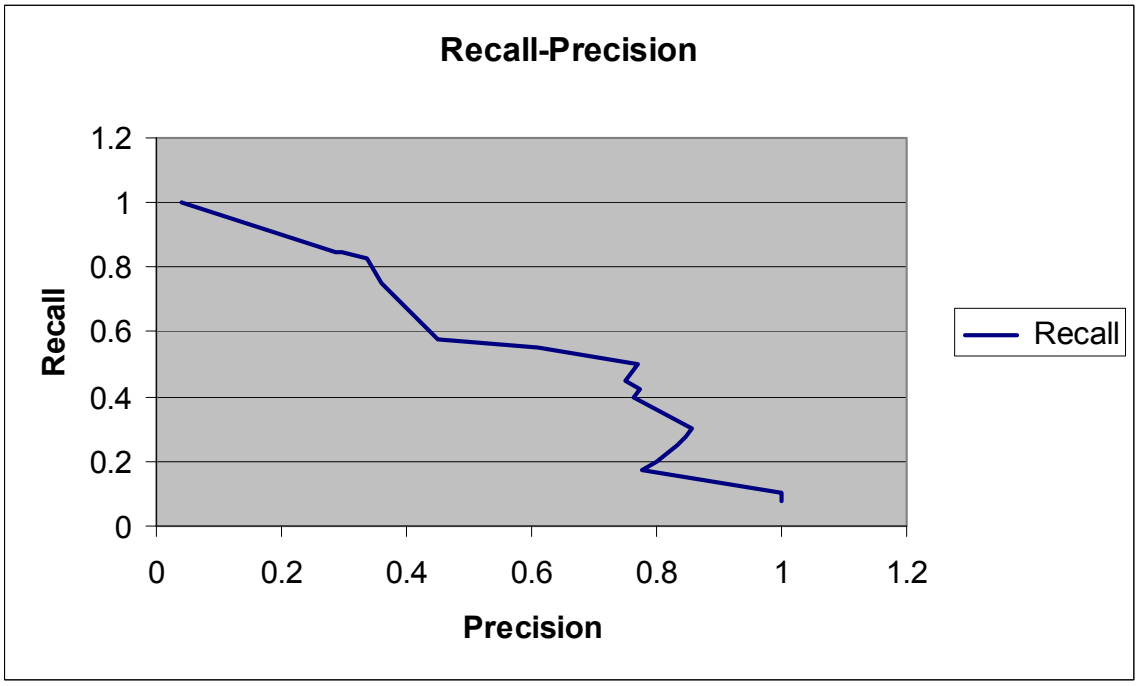

Figure 10. Overall recall-precision graph for TerrorPageCrawler.

The linguistic-pattern module provided additional filtering on sentences. We tested the system without the linguistic-pattern module (Figure 11), and performance was around 15% worse in mid-range and 25% worse at high recall. So linguistic-pattern information is definitely helpful even when good word clues are available. This is because there are relevant sentences for which the pattern provides the only evidence.

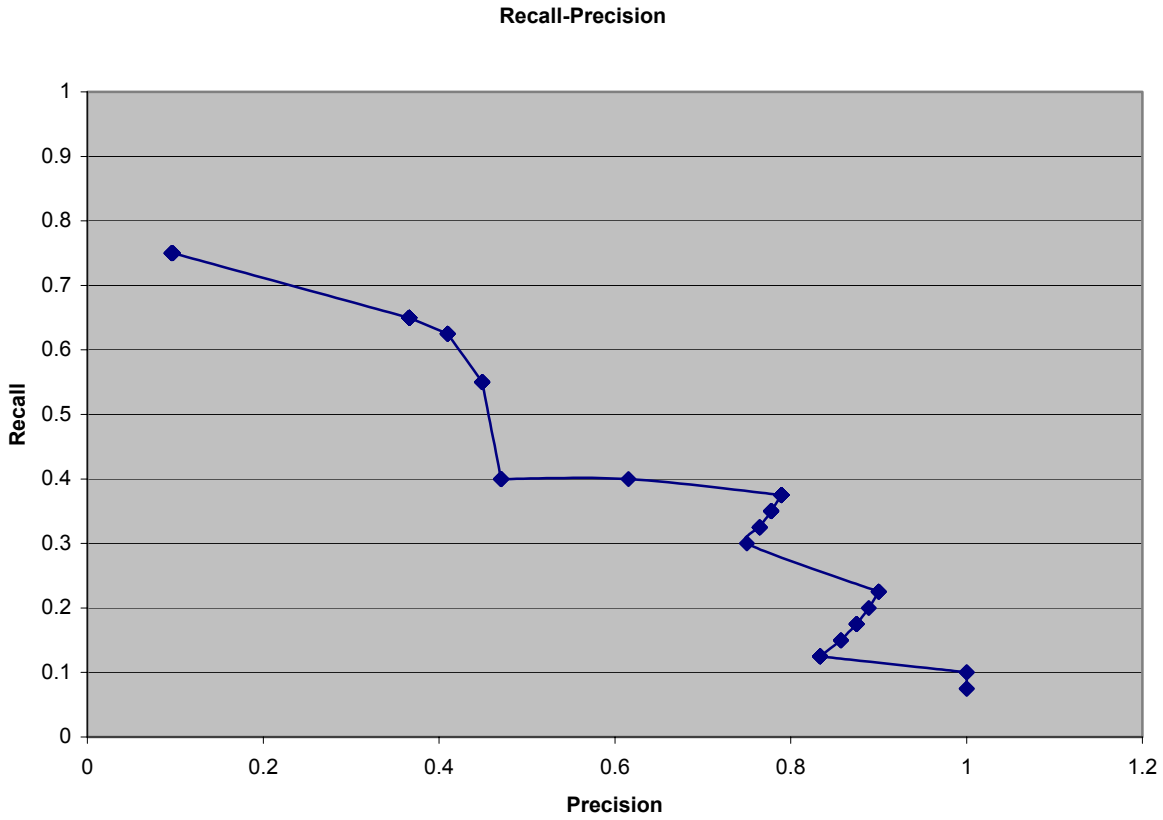

Figure 11: Recall-precision curve without the linguistic-pattern module.

For final tests, we set TerrorPageCrawler to start executing with a queue of 800 URLs. The crawler took approximately 16 hours to visit 5,500 pages running 4 threads. From these pages, 260,000 sentences were evaluated to have a probability of being terrorism-related of greater than 0.1. The linguistic-pattern matcher then reduced the number of sentences from 260,000 to 1,500 which were outputted to the database.

Accuracy of our search seemed to be high. For instance, a search of the database for the terrorist term "cell," a term not considered a strong clue by the learning process, did return a set of sentences containing useful information about terrorism (Figure 12). Other search terms such as "bomb" revealed many false alarms such as accounts of warfare, so further linguistic analysis would be necessary to improve the precision of the results.

Figure 12: Search results for the term "cell".

3. Conclusions

Perfect performance is impossible for data mining of intelligence information from unrestricted text. Our goal must be to yield as rich a source of data for human searching as possible. Our location-expression extractor program demonstrated that considerable summarization and categorization of information can be achieved automatically by a relatively small set of rules. While it had both false alarms and false negatives, it appears that the limited number of location-expression formats makes automation especially successful for this task. But there are additional issues in how to use the resulting information, which often is fuzzy and has time-development implications for a battle plan.

As for the terrorist-report program, it also yielded a rich set of data from a broad set of clues. This was richer than that found by a standard commercial browser, Alta Vista, since much unrelated information needed to be discarded from Alta Vista's candidate URLs. But the 1500 sentences matched by the post-processing utility represented 0.05% of the total sentences seen by the crawler, so we did not have statistically adequate data to find all the good word and phrase clues, and we need to conduct more experiments. In addition, performance could be improved by examining more non-word indicators of terrorism-related sentences.

The proper-noun module needs to be able to learn new proper nouns, since new person names and organizations appear in the news all the time. This could be done by distinguishing the sentence patterns in which unknown words occur, and using knowledge such as that certain verbs require people as agents. The proper-noun module also needs to recognize common misspellings such as Usama Bin Laden, Osama Bin Ladin, Al Qaeda, and Al Qaida. The linguistic-pattern matcher could be improved by preprocessing with a parts-of-speech tagger to rule out many obviously poor matches. It would also help to implement more word categories such as <place>, <time>, and <terrorist-act>, and to learn sentence patterns from experience by noting new patterns in which terrorism-associated words occur. The supervisor program could try to relate information about terrorist acts, which could lead to more successful recognition as in (Rubin, 2003). By recognizing the pieces of a terrorist plan, we can be more effective in interpreting related but ambiguous events ("connecting the dots").

Both projects illustrate that XML is not a necessity to collect and interpret intelligence. While forcing everyone to enter their data into a form and converting it to XML would simplify analysis, this can be a big burden on the author of an intelligence report, whose most interesting intelligence often are ideas that do not fit well into a form, and is simply not possible with the free-form prose of news reports from the media.

4. References

Baeza-Yates, R. (1998). Searching the Web: challenges and partial solutions. *Proc. String Processing and Information Retrieval: A South American Symposium*, 23–31.

Barcala, F., Vilares, J., Alonso, M., Grana, J., & Vilares, M. (2002, September). Tokenization and proper noun recognition for information retrieval. *Proc. 13th Intl. Workshop on Database and Expert System Applications*, 246-250.

Brill, E. (1998, May). Machine learning and automatic linguistic analysis: the next step. *IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, Vol. 2, 1033-1036.

Custy, J., McDonnell, J., & Gizzi, N. (2002, June). Estimating position and motion of mobile targets. *Command and Control Research and Technology Symposium*, Monterey, CA.

Gruenwald, L., McNutt, G., & Mercier, A. (2003. September). Using an ontology to improve search in a terrorism database system. *Proc. 14th Intl. Workshop on Database and Expert System Applications*, 753-757.

Heaton, J. (2002). *Programming Bots, Spiders, and Aggregators in Java.* San Francisco, CA: Cybex.

Kim, J.-T., & Moldovna, D. (1995, October). Acquisition of linguistic patterns for knowledge-based information extraction. *IEEE Trans. on Knowledge and Data Engineering, 7*, 5, 713-724.

Kolari, P., & Joshi, A., (2004, July-August). Web mining: research and practice. *Computing in Science & Engineering, 6*, 4, 49-53.

McCurley, K.S.; Tomkins, A., (2004, May). Mining and knowledge discovery from the Web. *Proc. of Conference on Parallel Architectures, Algorithms and Networks*, 4-9.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K., (1990, Winter). Five papers on Wordnet. *International Journal of Lexicography 3*, 4.

Morimoto, Y., Aono, M., Houle, M, & McCurley, K. (2003, January). Extracting spatial knowledge from the Web. *Proc. Intl. Conf. on Applications and Internet*, 326-333.

Petasis, G., Cucchiarelli, A., Velardi, P., Paliouras, G., Karkaletsis, G., & Spyropolous, C. (2000). Automatic adaptation of proper noun dictionaries through cooperation of machine learning and probabilistic methods. *Proc. 23rd Annual ACM Conf. on Research and Development in Information Retrieval*, Athens, Greece, 1228-1235.

Popp, R., Armour, T., Senator, T., & Numrych, K. (2004, March). Countering terrorism through information technology. *Communications of the ACM, 47*, 3, 36-43.
Rowe, N. (1997, June). Obtaining optimal mobile-robot paths with non-smooth anisotropic cost functions using qualitative-state reasoning. *International Journal of Robotics Research, 16*, 3, 375-399.
Rowe, N. (2002, July/August). MARIE-4: A high-recall, self-improving Web crawler that finds images using captions. *IEEE Intelligent Systems*, Vol. 17, no. 4, 8-14.
Rowe, N. (2004, June). Understanding Navy technical language via statistical parsing. *Command and Control Research and Technology Symposium*, San Diego, CA.
Rubin, S., Smith, M., & Trajkovic, L. (2003, October). A blackboard architecture for countering terrorism. *IEEE Intl. Conf. on Systems, Man, and Cybernetics*, San Diego, CA, vol. 2, 1550-1553.

Acknowledgements: This work was sponsored in part by DARPA as part of the TEMMPTS project of SPAWARSYSCEN, U.S. Navy, San Diego, CA. Views expressed are those of the authors and do not represent policy of the U.S. Navy.