1-1-2008

Natural Language Processing and e-Government: Crime Information Extraction from Heterogeneous Data Sources

Chih Hao Ku '12
Claremont Graduate University

Alicia Iriberri '06
Claremont Graduate University

Gondy Leroy
Claremont Graduate University

Natural Language Processing and e-Government: Crime Information Extraction from Heterogeneous Data Sources

Chih Hao Ku
Claremont Graduate University
School of Info. Systems & Tech.
130 E. Ninth St., Claremont, CA
Ku.justin@gmail.com

Alicia Iriberri
Claremont Graduate University
School of Info. Systems & Tech.
130 E. Ninth St., Claremont, CA
Alicia.IriberriAjuria@cgu.edu

Gondy Leroy, Ph.D.
Claremont Graduate University
School of Info. Systems & Tech.
130 E. Ninth St., Claremont, CA
Gondy.leroy@cgu.edu

ABSTRACT

Much information that could help solve and prevent crimes is never gathered because the reporting methods available to citizens and law enforcement personnel are not optimal. Detectives do not have sufficient time to interview crime victims and witnesses. Moreover, many victims and witnesses are too scared or embarrassed to report incidents. We are developing an interviewing system that will help collect such information. We report here on one component, the crime information extraction module, which uses natural language processing to extract crime information from police reports, newspaper articles, and victims' and witnesses' crime narratives. We tested our approach with two types of document: police and witness narrative reports. Our algorithms extract crime-related information, namely weapons, vehicles, time, people, clothes, and locations. We achieved high precision (96%) and recall (83%) for police narrative reports and comparable precision (93%) but somewhat lower recall (77%) for witness narrative reports. The difference in recall was significant at $p < .05$. We then used a spell checker to evaluate if this would help with witness narrative processing. We found that both precision (94 %) and recall (79%) improved slightly.

Categories and Subject Descriptors

H.3 [**Information Storage Retrieval**]: Information Extraction – *extract crime-associated information from heterogeneous data sources.*

H.3.1 [**Content Analysis and Indexing**]: Linguistic Processing – *analyze police and witness narratives.*

General Terms

Algorithms, Measurement, Performance, Design.

Keywords

Information Extraction. Natural Language Processing.

1. INTRODUCTION

In the United States, a property crime[1] is committed on average every 3 seconds; every 22 seconds, a violent crime[1] is committed. Today, most crimes are solved after investigators interview witnesses and victims, analyze the resulting narrative reports, and combine the information. Missing, nonresponse, or partial nonresponse data often occurs when data is collected using questionnaires or structured interviews [1]. Interviewers may fail to record data, want to finish an interview quickly, or record data incorrectly or illegibly [1]. In addition, analyzing salient clues within a significant number of narrative reports is an arduous task for police investigators and efficiency could be improved if text reports could be automatically analyzed to obtain relevant crime-related information, such as vehicle names, addresses, narcotics, and people's identities [2]. Information technology, e.g., information extraction, can be used to obtain crime-related information more efficiently.

We are developing a crime information extraction (IE) system to facilitate direct reporting by witnesses and victims. We use witness narrative reports obtained from online forums and blogs and police narrative reports from police departments and newspaper articles. We will integrate this IE system into our online crime reporting and interviewing system, which will guide and encourage witnesses and victims to report crime incidents so that police investigators can obtain more valuable information [3]. Automatically extracting information, combining information from crime narrative reports, and presenting a meaningful summary for police investigators will help them quickly comprehend crime incidents without having to read an entire report. Working with original witness crime reports that contain first-hand information is ideal, but difficult to achieve. In general these documents are difficult to obtain due to confidentiality or privacy concerns, and they often contain spelling and grammatical errors.

A successful and useful crime information extraction system should achieve high precision and recall regardless of the type and origin of the information. We complied a rich and substantial lexicon, by analyzing crime corpora from heterogeneous data sources. We developed crime information extraction rules that leverage the lexicons to extract useful information from various sources and with diverse formats. We compared performance of our IE for police narrative reports and witness or victim narrative reports. We report on the evaluation of the system by analyzing precision and recall.

2. INFORMATION EXTRACTION

Principles. Information extraction stands for a wide range of techniques that are applied to automatically extract pre-specified elements. For example, in biology, gene or protein names can be extracted from text [4]. In e-government, more specifically in crime reporting, the goal of IE techniques is to help investigators extract crime-related information quickly and effectively.

[1] FBI 2006 Crime Clock, http://www.fbi.gov/ucr/cius2006/about/crime_clock.html

Data/Text. Most existing crime information extraction projects use narrative reports coming primarily from police departments. For example, the Coplink project [5] used data from the Tucson and Phoenix police departments, the Uniform Crime Report (UCR) [6], which is report data collected by the Federal Bureau of Investigation (FBI), and from newspaper articles, newswire, and broadcast news included in the ACE and MUC corpora [7, 8]. They utilize a neural network to extract entities from the Tucson Police Department's narrative reports. Similarly, Lyons et al. [6] developed a web-based repository called GRASP to store UCR data. However, they allow users to clean data in three ways - automatic, manual, or a combination of both - so that missing, false, and fragmentary data can be removed. Feldman et al. [7] developed the TEG system combining knowledge engineering and machine learning approaches to evaluate MUC-7, ACE-2, and an industry corpus to extract entities.

Most research uses a single data source for named entity extraction [2, 9, 10]. Multiple data sources usually consist of different types of newspaper articles [7, 8]. All these sources comprise text that often has an explicit structure and from which typos have been removed. Narrative reports from police departments contain fixed data structures such as type of crime, date, time, and location. Newspaper articles are well organized and contain few typos and grammatical errors. Police narrative reports are often edited by police officers to remove grammatical and spelling errors to enhance readability and to protect people's privacy. As a consequence, high precision (75% to 98%) and recall (higher than 70%) for people's names, organization names, and narcotic drug names can be achieved. Relatively lower precision (below 70%) and recall (below 50%) for address and personal property were noted by Chau et al. [2] due to a variety of formats and abbreviations of addresses and a great diversity of personal property items.

Extracted Entities. Information extraction aims to extract specific, predefined entities from texts. Feldman and Sanger [11] point out that entities such as people, companies, locations, attributes (e.g., age of a person), facts (e.g., a relationship between two entities), and events (e.g., a terrorist act) can be extracted. Existing research has used people's names, locations, organizations' names [5], races, genders, age, weapons, crime types [12], addresses, narcotic drugs, vehicles, and personal properties [2] from crime narrative reports. According to Pentland [13], identifying salient clues underlying a pattern of incidents (e.g., sequence in time, focal actors, and indicators of content as well as context.) is important. Based on previous research, we decided to extract information related to names, pronominal nouns, times, vehicles, weapons, personal features, scenes, personal properties, colors, body parts, acts, events, and clothes.

Common Information Extraction Techniques. Chau et al. [2] specify techniques that can be used and combined for IE: lexical lookup [14], statistical methods, machine learning, and rule-based methods [15]. Most entity extraction systems match names of entities using lexical lookups. For example, Roark et al. [16] used noun-phrase co-occurrence to extract category entries from an online document. Machine learning [17] utilizes learning algorithms as opposed to hand-crafted rules to extract targeted information. For example, Chau et al. used neural networks [5] to extract named entities from police narrative reports and Liu et al. used Hidden Markov Models [18] to extract header information from computer science research papers. These approaches need large training data sets. In contrast, an advantage of rule-based methods, also pointed out by Maynard et al. [19], is that they can avoid ambiguity while not requiring a large training data set.

In addition to extracting individual entities, more complex structures can be identified. For example, Nath [12] utilized data mining to identify crime patterns. The author believes that clustering techniques can detect newer and unknown patterns better than supervised techniques such as classification. He used K-means clustering technique, which clusters objects based on attributes in his research.

Instead of using only one approach, most ongoing research usually integrates two or more approaches for IE extraction. For example, Chau et al. [2] used a neural network, lexical lookup, and a machine learning approach to extract name entities. In this paper, we adopt the rule-based and lexical lookup approaches.

3. CRIME INFORMATION EXTRACTION SYSTEM DEVELOPMENT

In the next sections, we illustrate how we developed the lexicon. We then explain which GATE components we employed and how they were modified for our goals. We also describe an additional algorithm that is needed to select the most relevant noun phrases.

3.1 Lexicon Development

To our knowledge, there is no crime ontology or vocabulary that is readily available to researchers. This is probably due to the wide range of terms needed such as weapons, vehicles, scenes, clothes, shoes, and physical features. We employed different strategies to collect lexicons for developing basic lexical lookup tables.

Data Sources. We collected data to develop the lexicons from five categories. The first category is official crime information Uniform Crime Reports (UCR) and the Federal Bureau of Investigation (FBI) web sites provide us with definitions of crimes. Lexicons containing crime types, e.g., robbery, and weapons, e.g., guns, can be collected from official police web sites. The second category is encyclopedia information from resources such as Wikipedia[2] and MSN Encarta[3]. When building the vehicle and weapon lexicons, Wikipedia and Encarta provided many main- and sub-categories of vehicle and weapon information. The third category contains general web sites and blogs. Different keyword combinations were used to search for useful web sites to complete the lexicons. For example, The Serious Wheels[4] web site contains almost every car brand with detailed information. The fourth category contains information from structured knowledge sources, such as FrameNet[5]. FrameNet is useful for collecting abstract lexicons such as scenes and physical features before searching for web sites and online encyclopedias. The last category contains thesauri and dictionaries. We utilized them to expand the lexical lookup set with additional information such as synonyms. The thesauri dictionaries we adopted include Thesaurus.com[6], Collins Cobuild, and MSN Encarta. The Collins Cobuild[7] dictionary incorporates rich informal English, slang, and a thesaurus. The most frequently

[2] Wikipedia, http://wikipedia.org/
[3] MSN Encarta, http://encarta.msn.com/
[4] Serious Wheels, http://www.seriouswheels.com/cars.htm
[5] FrameNet, http://framenet.icsi.berkeley.edu/
[6] Thesaurus.com, http://thesaurus.reference.com/
[7] Collins Cobuild, http://www.elearnaid.com/basiccobuildcd.html

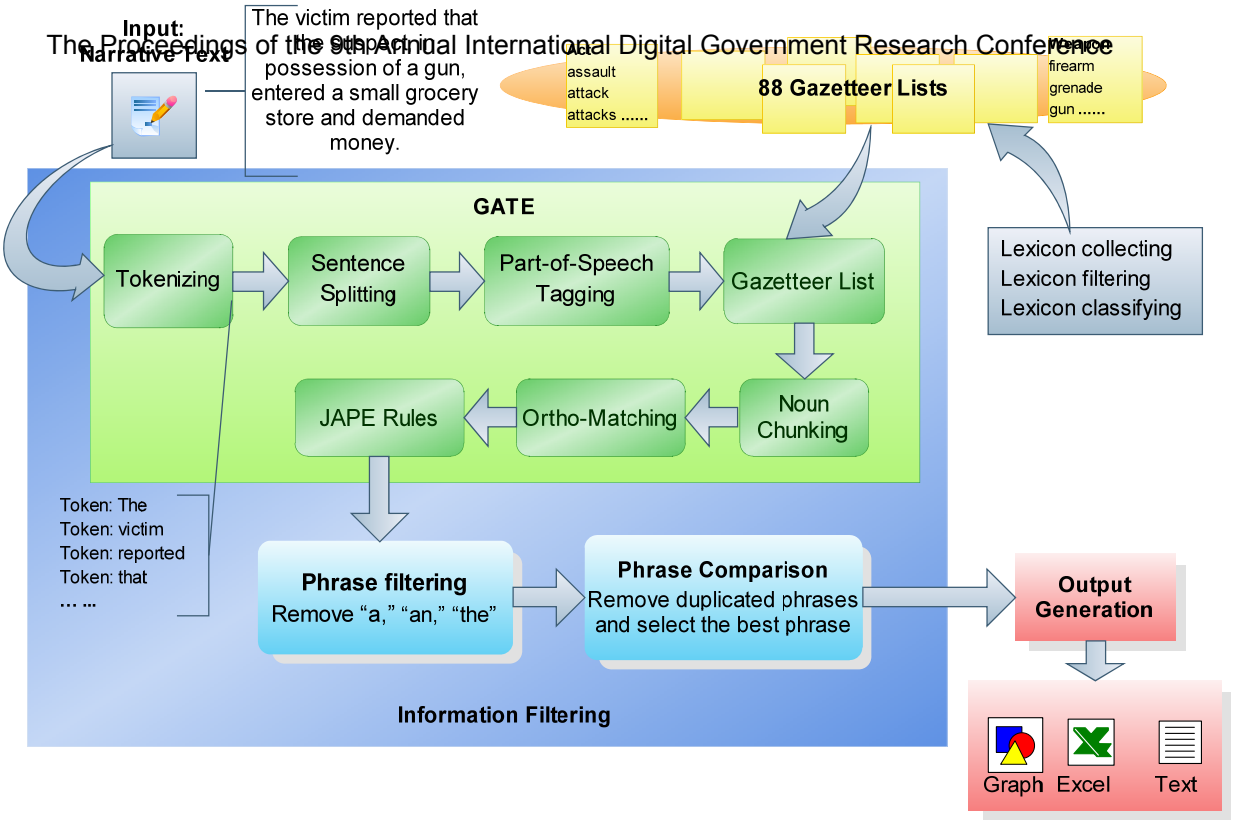

Figure 1. Crime Information Extraction Module

used definition of a word is listed as first and black diamonds indicate the frequency of word usage, as shown in Figure 2. Five black diamonds represent that a word is most frequently used; no diamond represents that a word is rarely used. In this example, the word "attack" is often used and the most common definition is to attack a person or place.

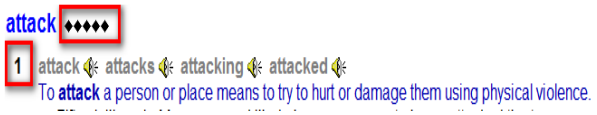

Figure 2. An example of the Collins Cobuild dictionary

Duplicate Filtering. Overall, the lexicon comprises eighty-eight gazetteer lists. A gazetteer contains a list of names stored in a plain text file. Each such gazetteer list contains a large number of words related to crime, for example, our gazetteers cover categories such as 'Act', 'Scene', 'People', 'Personal Property', 'Vehicle', 'Weapon', 'Body Part', 'Time', and 'Clothing'. Each category contains sub-categories. For example, the 'Clothing' category contains shoes and clothes.

To remove duplicates, we relied on word frequency to determine where to retain words. For example, the word "white" appears in race, car brand, and color. According to the Collins Cobuild dictionary, it rarely refers to a car brand and so we removed "white" from car brands. In some cases, duplicates cannot be removed and then we use rules to disambiguate the words, e.g., when "white" refers to race or a car color.

3.2 GATE Modules

We employed the General Architecture for Text Engineering (GATE) system [20] to build our IE modules. GATE is an open source framework [21] which is comprised of several modules. The modules we adopted (see Figure 1) include the tokenizer, gazetteer, sentence splitter, part-of-speech (POS) tagger, noun chunks, ortho-matcher, and JAPE rules. We explain each step in our algorithm using the example "The victim reported that the suspect, in possession of a gun, entered a small grocery store and demanded money."

Tokenizer. The tokenizer splits input text into tokens such as words, numbers, and symbols [10]. The narrative example (see Figure 1) is tokenized as:

The / victim / reported / that / the / suspect / , / in / possession / of / a / gun / , / entered / a / small / grocery / store / and / demanded / money /

Sentence Splitter. This component separates an input text into sentences. The example input only contains a single sentence. Thus, the result is:

/ The victim reported that the suspect, in possession of a gun, entered a small grocery store and demanded money./

POS Tagger. This is a revised version of the Brill tagger [22]. Every token is annotated with a POS (part-of-speech) tag, which is a grammatical tag such as noun, determiner, or adverb. There are 42 different possible tags as default in GATE. For example, in our sentence DT refers to determiner, NN refers to singular or mass nouns, VBD refers to past tense verbs, JJ refers to adjectives, CC refers to coordinating conjunction, and IN refers to prepositions. Thus, the result is:

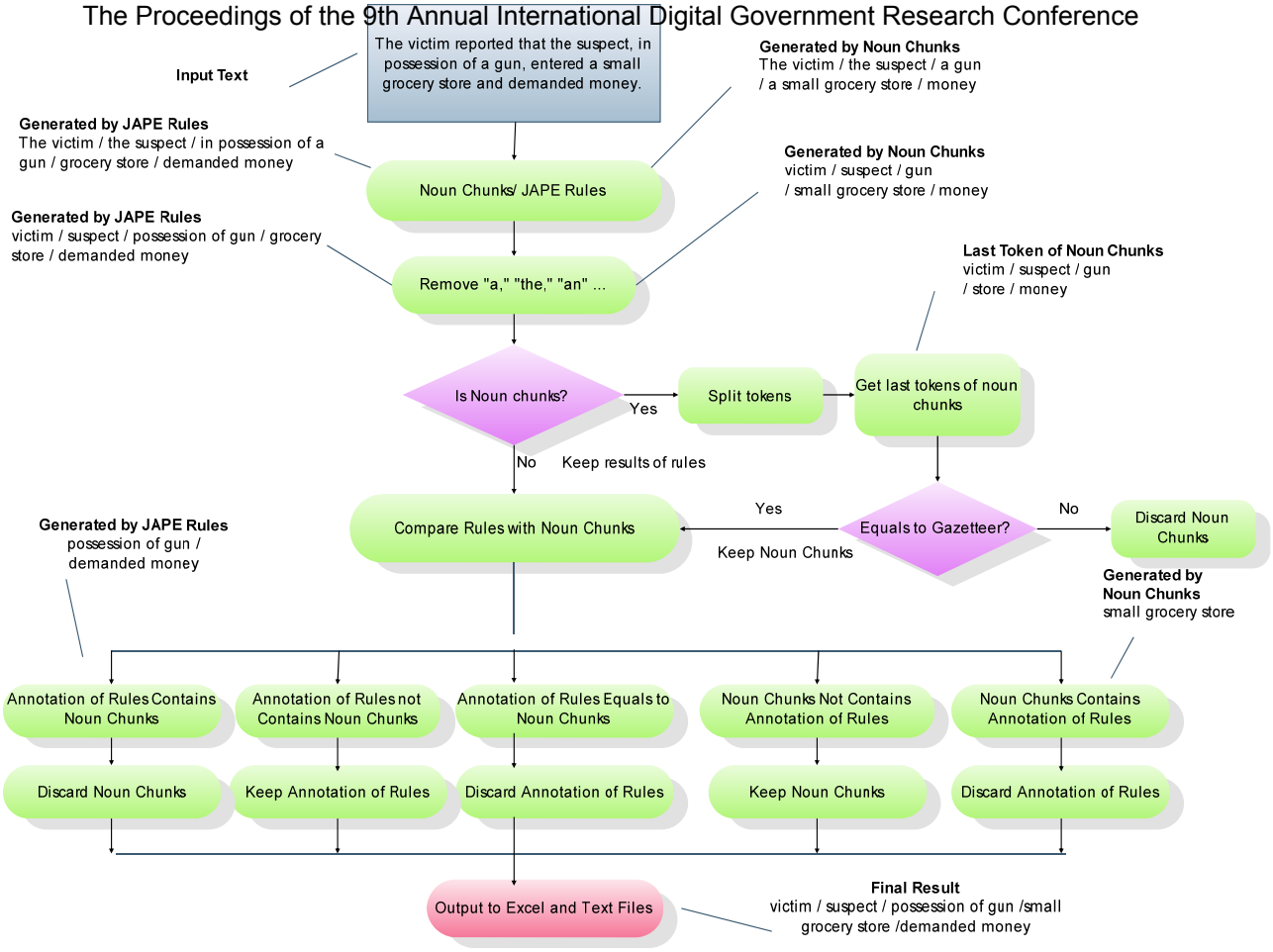

Figure 3. Phrase Filtering and Comparison Algorithm

*The***(DT)** */ victim***(NN)** */ reported***(VBD)** */ that* **(IN)** */ the***(IN)** */ suspect***(NN)** */ ,***(,)** */ in***(IN)**/ *possession***(NN)** */ of***(IN)** */ a***(DT)** */ gun***(NN)** */ ,***(,)** */ entered***(VBD)** */ a***(DT)** */ small***(JJ)** */ grocery***(NN)** */ store* **(NN)** */ and***(CC)** */ demanded***(VBD)** */ money***(NN)** */

Noun Phrase Chunker. This component marks noun phrases in input texts [23]. The result generated by the noun phrase chunker is:

{The victim} {the suspect} {possession} {a gun} {a small grocery store} {money}

Gazetteer. This component is mainly used to identify entities (e.g., act and event) and can also be used in rules to extract complex phrases. We modified this component and we use 88 gazetteers (described above). An example of the gazetteer list for 'Act' is:

Assault/ attack / fight / kill / massacre / murder / rape / shoot / slay / stab / torture / steal / rob / burgle / plunder / ransack / grab / …

Ortho-Matcher. This component recognizes specific names such as cities, streets, and states. These names may contain upper initials, all capitals, or all lower case letters. For example, the *{Token.orth == upperInitial}* section can be used to recognize the entity "California".

JAPE Rule. JAPE (Java Annotations Pattern Engine) language [20] is used for pattern matching and to add annotations to the matched patterns. We created fourteen JAPE files. Each JAPE file combines several macros and rules. For example, the StreetMacro is used to recognize street addresses in text.

```
Macro: StreetMacro
(
 (Number)(Word)(StreetLookup)(cardinalDirection)
)
```

This *StreetMacro* encompasses another four macros, namely *Number*, *Word*, *StreetLookup*, and *cardinalDirection*. An address such as "53rd Ave N." can be extracted from this macro. "53" refers to *Number*, "rd" refers to *Word*, "Ave" refers to *StreetLookup*, and "N." refers to *cardinalDirection*.

The strength of JAPE rules is that the macros and rules can integrate and combine POS and gazetteer list tags to extract meaningful phrases. For example, the phrase "small grocery store" has a richer semantic meaning than the word "store" so police investigators know more about the crime scene. JAPE rules can also incorporate results from rules applied earlier. An example rule for vehicle type is:

```
Rule: VehicleTypes
        Priority: 112
          (
           (VehicleDashRule)
```

```
    | {Lookup.minorType == vehicleType}
   )
  :vehicle -->
  :vehicle.Rule = { majorType= "vehicle",
  minorType="vehicleType", Rule = "Vehicle_Rule"}
```

The *{Lookup.minorType == vehicleType}* section extracts all types of vehicles stored in our gazetteer lists and the *(VehicleDashRule)* section extracts phrases such as 4-door MAZDA3, which requires a rule due to the hyphen between "4" and "door." We manipulated the sequence of rules using the priority attribute since some rules are more important than others. This also makes it possible to use clean-up rules to remove duplicate or undesired results.

3.3 Crime Information Extraction Algorithm

The noun phrases generated by the noun phrase chunker and retained in the previous phases contain noisy information, such as "information" and "knowledge," which are common but not very relevant words in our context. Moreover, the noun phrase chunker may generate the same noun phrases that the JAPE rules generate. In some cases the noun chunker's output is better; in other cases the JAPE rules output is better.

To select the best and most meaningful phrases, we define a meaningful phrase as a longer phrase. Usually, a long phrase is better than a short phrase or a word because it contains more details useful to e.g., detectives. For example, "9th and Spring Street" is better than "street" since the first phrase contains more detailed information about the location. We used this heuristic to address duplicate nouns with an additional filtering algorithm.

We use the same example to illustrate the filtering algorithm (see Figure 3). The first step is to remove determiners, such as "a," "an," and "the," prepositions at the beginning of phrases, and words on our stopword lists, e.g., "all" and "but," from phrases generated by the JAPE rules or the noun phrase chunker. For example, the phrase "a small grocery store" will become "small grocery store." Next, the algorithm detects if a phrase is generated by the JAPE rules or the noun phrase chunker.

If the phrase is generated by the noun phrase chunker, the system runs another tokenizer to obtain the last token in the noun phrase. For example, the last token in "small grocery store" is the word "store." After that, the system compares this last token to gazetteer lists and discards unwanted noun phrases. For example, the phrase "political history" has nothing to do with crimes and the word "history" is not contained in our gazetteer lists. Consequently, the phrase "political history" will be discarded. After this step, every noun phrase is considered to be related to crime.

If a phrases is generated by both the JAPE rules and the noun phrase chunker, we consider the longer phrase to be more relevant and useful. For example, "school bus" is more useful than "bus" since it contains more detailed information. For this reason, the system compares the length of both phrases. If the first phrase contains the second phrase, then the system will keep the first phrase and discard the second phrase and vice versa. For example, the phrase "small grocery store" contains the phrase "grocery store" so the system will keep the first phrase and discard the second phrase. If the first phrase is equal to the second phrase, the system will discard the phrase generated by the JAPE rules. If the phrases generated by the JAPE rules and by the noun phrase chunker do not contain each other, the system retains both of them. The result for our example text (see Figure 1 input text) is:

victim / suspect / possession of gun / small grocery store / demanded money

Currently, the system can output results to an Excel and a text file for further evaluation. The extracted entities are also stored in our database to ask follow up questions according to the principles of cognitive interview. The cognitive interview [24] is a psychological technique that helps people reminisce more information about an incident.

4. EVALUATION

We collected a crime report corpus from heterogeneous data sources for our system development and evaluation. The corpus used for evaluation was never used during system development.

4.1 Crime Narrative Sources

Unsolved-Crimes. Unsolved-Crimes International[8] is a volunteer organization and provides violent crime information from its own web site and a Yahoo! group. The information includes news articles, police reports, and witness narrative reports.

SUBA District Unit. SUBA[9] is a Police Officer team and provides police reports (2005-2007) from its own web site and a Google group.

Chat LawInfo. Chat LawInfo[10] is a forum, which contains first-hand narrative reports from witnesses or victims.

True Crime Blog. True Crime Blog[11] provides rich crime information mainly from news articles and books.

Baltimore Crime. Baltimore Crime[12] contains news and police reports and is updated frequently.

ExpertLaw[13]. This web page contains first-hand witnesses and victims' reports.

4.2 Methodology

Twenty police narrative reports and twenty witness narrative reports were randomly selected from our corpus. For each, we established a gold standard. One author created the gold standard for each document. She marked phrases according to pre-defined groups such as weapons, people, and vehicles and calculated the total number of entities in each group that should be extracted by the system. The results from the system are compared against this gold standard. We use both precision and recall to evaluate our approach. Recall is the ratio of the correctly extracted features to all of the correct features in the document. Precision is the ratio of the correctly extracted features to the total extracted features.

$$\text{Recall} = \frac{\text{Correctly Extracted Features}}{\text{All Correct Features in Document}}$$

$$\text{Precision} = \frac{\text{Correctly Extracted Features}}{\text{Total Extracted Features}}$$

[8] Unsolved-Crimes, http://groups.yahoo.com/group/Unsolved-Crimes/
[9] SUBA, http://groups.google.com/group/SPD_SDU/web/crime-bulletins
[10] Chat LawInfo, http://chat.lawinfo.com/
[11] True Crime Blog, http://laurajames.typepad.com/clews/
[12] Baltimore Crime, http://baltimorecrime.blogspot.com/
[13] ExpertLaw, http://www.expertlaw.com/forums/index.php

Table 1. Police Narrative Reports by Groups

Group	CE	TE	Precision	Recall
Act	51	51	100%	70%
Age	*3	*3	100%	75%
Body Part	9	9	100%	100%
Clothes	6	7	86%	86%
Drug	*1	*1	100%	50%
Electronic	*2	*2	100%	100%
Event	14	14	100%	67%
Face	10	10	100%	100%
Hair	*2	*2	100%	100%
Job	*NA	*NA	NA	NA
People	210	216	97%	92%
Physical Condition	*1	*1	100%	11%
Physical Feature	8	8	100%	89%
Property	29	29	100%	74%
Scene	101	108	94%	85%
Shoes	*NA	*NA	NA	NA
Time	43	49	88%	82%
Vehicle	15	17	88%	79%
Weapon	*2	*2	100%	67%
Total	507	529	96%	83%

CE = Correctly extracted items
TE = Total extracted items
* Refers to extracted fewer than 5 entities (not discuss in the paper)

Table 2. Witness Narratives by Groups

Group	CE	TE	Precision	Recall
Act	37	37	100%	55%
Age	10	13	77%	56%
Body Part	*NA	*NA	NA	NA
Clothes	*1	*1	100%	100%
Drug	6	6	100%	100%
Electronic	*2	*2	100%	100%
Event	6	6	100%	43%
Face	*4	*4	100%	100%
Hair	*2	*2	100%	100%
Job	*0	1	0%	0%
People	338	345	98%	91%
Physical Condition	*1	*1	100%	100%
Physical Feature	*0	*1	NA	0%
Property	29	29	100%	67%
Scene	50	74	68%	58%
Shoes	*1	*1	100%	100%
Time	20	24	83%	56%
Vehicle	28	29	97%	78%
Weapon	*1	*1	100%	50%
Total	536	577	93%	77%

CE = Correctly extracted items
TE = Total extracted items
* Refers to extracted fewer than 5 entities (not discuss in the paper)

4.3 Results

The average length of the 20 police narratives is 100 words while that of the 20 witness narratives is 150 words. Among those 20 witness narratives, the average number of typos and errors is 6 to 7 words. Common typos include "childs," "vechile," "illgal," "jewelery," and "burlgar" while common errors include "out side," "mexico," and "couldnt."

Overall we achieved satisfactory precision of 96% and 93% and recall of 83% and 77% for the police and witness narrative reports. See Table 1 and Table 2 for details.

For the police narratives, the system extracted fewer than 5 entities for the categories such as 'Age', 'Drug', and 'Weapon' and so we do not discuss them here. Overall precision was 96% and overall recall was 83%. We achieved 100% precision with 'Act', 'Body Part', 'Event', 'Face', 'Physical Feature', and 'Personal Property' and the highest recall for 'Body Part' and 'Face'. 'People' and 'Scene' were extracted most often from our collected narratives. We encountered the lowest precision for 'Clothes' and lowest recall for 'Event'. We achieved high precision ranging from 94% to 97% and recall ranging from 85% to 92% for both 'Scene' and 'People'. For 'Time', we revised existing JAPE rules of GATE and achieved satisfactory precision of 88% and recall of 82%.

For the witness narratives, the system extracted fewer than 5 entities for categories such as 'Clothes', 'Electronic', 'Physical Condition', and 'Weapon' and so we do not discuss them here. Overall precision was 93% and overall recall was 77%. We achieved 100% precision with the categories 'Act', 'Drug', 'Event', and 'Personal Property' and the highest recall (100%) for 'Drug'. We encountered the lowest precision for 'Scene' and lowest recall for 'Act'. We encountered low recall (below 60%) for 'Act', 'Age', 'Event', 'Scene', and 'Time'.

We compared precision and recall for both types of documents and found that the difference in recall was significant at $p < .05$ based on an independent samples t-test. The system achieved comparable precision for both types of document.

4.4. Impact of Spell Checking

Table 3. Witness Narratives by Groups
(With spell checker – select the first alternative)

Group	CE	TE	Precision	Recall
Act	37	38	97%	55%
Age	10	13	77%	56%
Clothes	*1	*1	100%	100%
Drug	6	6	100%	100%
Electronic	*2	*2	100%	100%
Event	7	7	100%	50%
Face	*4	*4	100%	100%
Hair	*2	*2	100%	100%
Job	*0	*1	0%	0%
People	339	345	98%	92%
Physical Condition	*1	*1	100%	100%
Physical Feature	*0	*1	NA	00%
Property	29	29	100%	67%
Scene	51	77	66%	57%
Shoes	*1	*1	100%	100%
Time	20	23	87%	57%
Vehicle	28	29	97%	78%
Weapon	*1	*1	100%	50%
Total	539	581	93%	78%

CE = Correctly extracted items
TE = Total extracted items
* Refers to extracted fewer than 5 times (not discuss in the paper)

Table 4. Witness Narratives by Groups
(With spell checker – select the best one)

Group	CE	TE	Precision	Recall
Act	39	39	100%	58%
Age	10	13	77%	56%
Clothes	*1	*1	100%	100%
Drug	6	6	100%	100%
Electronic	*2	*2	100%	100%
Event	8	8	100%	57%
Face	*4	*4	100%	100%
Hair	*2	*2	100%	100%
Job	*0	*1	0%	0%
People	338	345	98%	91%
Physical Condition	*1	*1	100%	100%
Physical Feature	*0	*1	NA	0%
Property	29	29	100%	67%
Scene	55	76	73%	63%
Shoes	*1	*1	100%	100%
Time	20	24	83%	56%
Vehicle	33	33	100%	92%
Weapon	*1	*1	100%	50%
Total	550	587	94%	79%

CE = Correctly extracted items
TE = Total extracted items
* Refers to extracted fewer than 5 times (not discuss in the paper)

To evaluate the impact of typos and other errors on our system, we submitted all twenty witness reports a second time after applying a spell checker. There was no need to do this for the police narratives, since they do not contain spelling errors.

We utilized the open-source spell checker called Ekit[14] to conduct the second round of evaluation. We evaluated different approaches to correct spelling errors. The first method is to select the first alternative word to correct errors. This would be easy to automate. The second method is that a person (ideally the writer of the narrative) selects the best alternative word. A spell check would then need to be included in the interviewing system. For each approach, we calculated precision and recall after correcting errors.

For the witness narratives with the first alternative word, overall precision was 93% and overall recall was 78%. See Table 3 for details. We achieved 100% precision with the categories 'Drug', 'Event', and 'Personal Property' and the highest recall for 'People'. We encountered the lowest precision for 'Scene' and lowest recall for 'Act'. The overall result shows that the precision for 'Act' (97%) is slightly lower than the original one without the spell checker. For 'Time', precision (87%) was slightly improved and for 'Event', recall (50%) was slightly improved.

For the witness narratives with the best alternative word, overall precision was 94% and overall recall was 79%. See Table 4 for details. We achieved 100% precision with the categories 'Act', 'Drug', 'Event', 'Vehicle', and 'Personal Property' and the highest recall for 'Vehicle'. We encountered the lowest precision for 'Scene' and lowest recall for 'Age' and 'Time'. We noticed that precision was improved for 'Scene' and 'Time' while recall was improved for 'Act', 'Event', 'Scene', and 'Vehicle'.
Comparing the resulting precision and recall after spelling correction with witness reports shows that the difference in recall is significant ($p < .05$) with the first alternative word correction method, but that precision and recall are equal when the best word is selected for correction.

5. DISCUSSION

The majority of our errors were related to the categories 'Act' and 'Event', which are easily confused and resulted in lower recall for both police and witness narratives. The main distinction between 'Act' and 'Event' is that 'Act' contains verbs and 'Event' contains

[14] Ekit, (http://www.hexidec.com/ekit.php)

nouns. For example, the 'Act' list contained attack, destroy, fight, and kill. We will revisit our lists and filter words that are not important enough to be included in our gazetteer lists The low recall can be further improved by expanding gazetteer lists of 'Act' and 'Event'.

For the witness narratives, the system missed a few important phrases such as "three-week period" and "two weeks" and extracted the number in front of a street as a year. For instance, the system extracted the number "2500" as a year from the address "2500 block of Cross Hill Court". Such problems result in low recall for 'Scene' and 'Time'. These can be resolved by improving our rules.

To assess the usefulness of a spell checker, we compare the overall precision and recall for three types of witness narratives. The overall result indicates that selecting the best alternative words improved precision and recall for 'Act', 'Event', 'Scene', and 'Vehicle'.

6. CONCLUSION and FUTURE WORK

An online reporting system with IE technology allows people who are embarrassed or scared to report crimes anonymously. It can also be useful to complement police interviews with known victims or witnesses. We are developing an IE system to collect relevant crime information to help police investigators solve crimes efficiently. The system leverages GATE components to extract crime-related information from police narratives and witness narratives. We evaluated the IE module with 20 representative police and 20 witness narratives. The IE module achieved high precision and recall for police narratives but slightly lower precision and recall for witness narratives. We also evaluated the usefulness of a spell checker. The overall results indicate that selection of the best alternative word by a human user is the best way to enhance both precision and recall.

We are combining the IE system with principles of the cognitive interview [3]. Our extracted entities are used to generate follow-up questions. We encourage the use of natural language to report crimes so people are not required to fill out numerous structured reports or questionnaires. Our goal is to provide a reliable online crime reporting system that people can report crime incidents anonymously. We hope to expand and improve our approach by integrating pronominal resolution and relationship assignments. Both will allow us to generate a graph, which explains relationships between entities and may serve as a visual summary. For example, a suspect may have a relation with a weapon (i.e., be the owner).

7. REFERENCES

[1] E. D. d. Leeuw, "Reducing Missing Data in Surveys: An Overview of Methods," *Quality and Quantity*, vol. 35, pp. 147-160, 2001.

[2] M. Chau, J. J. Xu, and H. Chen, "Extracting Meaningful Entities from Police Narrative Reports," in *ACM International Conference Proceeding Series*. vol. 129 Los Angeles, California: Digital Government Research Center, 2002, pp. 1-5.

[3] A. Iriberri and G. Leroy, "Natural Language Processing and e-Government: Extracting Reusable Crime Report Information," in *Information Reuse and Integration, 2007. IRI 2007. IEEE International Conference on*, Las Vegas, NV, USA, 2007, pp. 221-226.

[4] L. Hirschman, A. Yeh, C. Blaschke, and A. Valencia, "Overview of BioCreAtIvE: Critical Assessment of Information Extraction for Biology," *BMC Bioinformatics*, vol. 6, May 2005.

[5] H. Chen, W. Chung, J. J. Xu, G. Wang, Y. Qin, and M. Chau, "Crime Data Mining: A General Framework and Some Examples," *Computer*, vol. 37, pp. 50-56, April 2004.

[6] A. M. Lyons, J. Michael W. Packer, M. B. Thomason, J. C. Wesley, P. J. Hansen, J. H. Conklin, and D. E. Brown, "Uniform Crime Report "SuperClean" Data Cleaning Tool," *Systems and Information Engineering Design Symposium, 2006 IEEE*, pp. 14-18, 2006.

[7] R. Feldman, B. Rosenfeld, and M. Fresko, "TEG - A Hybrid Approach to Information Extraction," *Knowledge and Information Systems*, vol. 9, pp. 1-18, Jan. 2006.

[8] D. Maynard, K. Bontcheva, and H. Cunningham, "Towards a Semantic Extraction of Named Entities," in *Recent Advances in Natural Language Processing* Bulgaria, 2003.

[9] K. Pastra, D. Maynard, O. Hamza, H. Cunningham, and YorickWilks, "How Feasible is the Reuse of Grammars for Named Entity Recognition?," in *In Proceedings of the Language Resources and Evaluation Conference*, 2002, pp. 1412-1418.

[10] K. Bontcheva, M. Dimitrov, D. Maynard, V. Tablan, and H. Cunningham, "Shallow Methods for Named Entity Coreference Resolution," in *Workshop TALN 2002* Nancy, France, 2002.

[11] R. Feldman and J. Sanger, "Information Extraction," in *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*: Cambridge University Press 2006, pp. 94-130.

[12] S. V. Nath, "Crime Pattern Detection Using Data Mining," in *2006 IEEE/WIC/ACM International Conference on*, 2006, pp. 41-44.

[13] B. T. Pentland, "Building Process Theory with Narrative: From Description to Explanation," *Academy of Management Review*, vol. 24, pp. 711-724, Oct. 1999.

[14] A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman, "NYU: Description of the MENE Named Entity System as Used in MUC-7," in *in Proceedings of the Seventh Message Understanding Conference (MUC-7)*, April 1998.

[15] M. Maslennikov, H.-K. Goh, and T.-S. Chua, "ARE: Instance Splitting Strategies for Dependency Relation-based Information Extraction," in *Proceedings of the COLING/ACL on Main conference poster sessions*, Sydney, Australia, 2006, pp. 571-578.

[16] B. Roark and E. Charniak, "Noun-Phrase Co-Occurrence Statistics for Semi-Automatic Semantic Lexicon Construction," in *Proceedings of the 17th international conference on Computational linguistics* Montreal, Quebec, Canada: Association for Computational Linguistics, 1998, pp. 1110-1116.

[17] D. Freitag, "Machine Learning for Information Extraction in Informal Domains," *Machine Learning*, vol. 39, pp. 169 - 202, Nov. 2000.

[18] Y. Liu, Y. Lin, and Z. Chen, "Using Hidden Markov Model for Information Extraction Based on Multiple Templates," in *Natural Language Processing and Knowledge Engineering, 2003. Proceedings*, 2003, pp. 394- 399.

[19] D. Maynard, M. Yankova, A. Kourakis, and A. Kokossis, *Ontology-Based Information Extraction For Market Monitoring And Technology Watch*. Heraklion, Crete, 2005.

[20] H. Cunningham, "GATE, a General Architecture for Text Engineering," *Computers and the Humanities*, vol. 36, pp. 223-254, May 2002.

[21] D. Maynard, K. Bontcheva, H. Saggion, H. Cunningham, and O. Hamza, "Using a Text Engineering Framework to Build an Extendable and Portable IE-based Summarisation System," in *Proceedings of the ACL-02 Workshop on Automatic Summarization* Phildadelphia, Pennsylvania: Association for Computational Linguistics, 2002.

[22] M. Hepple, "Independence and Commitment: Assumptions for Rapid Training and Execution of Rule-based POS Taggers," in *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics* Hong Kong: Association for Computational Linguistics, 2000.

[23] L. A. Ramshaw and M. P. Marcus, "Text Chunking using Transformation-Based Learning," in *Proceedings of the Third Workshop on Very Large Corpora*, 1995, pp. 82-94.

[24] A. Memon and R. Bull, "The Cognitive Interview - Its Origins, Empirical Support, Evaluation and Practical Implications," *Journal Of Community & Applied Social Psychology*, vol. 1, pp. 291-307, 1991.