Crime Information Extraction from Police and Witness Narrative Reports

Chih Hao Ku ’12
Claremont Graduate University

Alicia Iriberry ’06
Claremont Graduate University

Gondy A. Leroy
Claremont Graduate University

Recommended Citation
Crime Information Extraction from Police and Witness Narrative Reports

Chih Hao Ku, Alicia Iriberri, Gondy Leroy
School of Information Systems and Technology
Claremont Graduate University

ku.justin@gmail.com, Alicia.IriberriAjuria@cgu.edu, gondy.leroy@cgu.edu

Abstract—To solve crimes, investigators often rely on interviews with witnesses, victims, or criminals themselves. The interviews are transcribed and the pertinent data is contained in narrative form. To solve one crime, investigators may need to interview multiple people and then analyze the narrative reports. There are several difficulties with this process: interviewing people is time consuming, the interviews – sometimes conducted by multiple officers – need to be combined, and the resulting information may still be incomplete. For example, victims or witnesses are often too scared or embarrassed to report or prefer to remain anonymous. We are developing an online reporting system that combines natural language processing with insights from the cognitive interview approach to obtain more information from witnesses and victims. We report here on information extraction from police and witness narratives. We achieved high precision, 94% and 96%, and recall, 85% and 90%, for both narrative types.

1. INTRODUCTION
Homeland security focuses on the protection of populations and essential infrastructure [1]. Information technology can contribute by helping solve and prevent crimes more efficiently. In the United States, millions of crimes go unreported [2]. Many victims and witnesses are too scared or embarrassed to report crime incidents. In some cases, interviewers may record answers inaccurately or illegibly, or may fail to record them [3]. In addition, report data such as Uniform Crime Reports (UCR) collected by the FBI often contains missing, incomplete, or incorrect data [4].

We are developing an online reporting system that addresses these problems. It will allow those too scared or embarrassed to report anonymously and, because it is based on principles of the cognitive interview [5], it may also help them remember more information correctly. Our approach will also enable us to combine the reported information in one report for law enforcement personnel.

Our system combines information extraction (IE) and principles of the cognitive interview [6]. The cognitive interview is a psychological technique that helps people remember more information about an incident. In order to leverage its principles, we encourage the use of natural language so that people do not need to fill out numerous structured reports. Such forms may be complex or difficult to understand leading fields left blank and incomplete information. By using natural language, people can report a crime more easily and thus more information can be collected. To enable such reporting, we need to extract crime-relevant information to ask follow up questions and compile a final report. Our goal is to obtain as much information as possible. To this end, we developed a large lexicon that combined with rule-based system can extract crime-related entities. Those extracted entities are triggers for our system to ask questions according to the principles of the cognitive interview.

To evaluate the information extraction component, we collected police and witness narratives from web sites, blogs, and forums. We test these by calculating precision and recall of information in comparison with a separately developed gold standard. The diversity of data sources shows the dependability.

2. INFORMATION EXTRACTION
Information extraction aims to extract pre-specified elements. For example, the names of people, places, or organizations can be extracted from documents without “deep understanding” [7] of the text. IE techniques have been used for many different purposes such as to extract auction prices from eBay and Yahoo web pages [8], to extract text information from PDF files [9], or in bioinformatics, to extract named entities and relationships of genes, proteins, and RNA from scientific publications [10].

Commonly extracted crime-related entities are race, gender, age, weapons [11], addresses, narcotic drugs, vehicles, and personal properties [12]. However, such annotated
Unfortunately, there are also several disadvantages. The cognitive interview has many advantages. For example, Chau and others [12, 15] proposed a system that used noun phrasing, lexical lookup, and neural network to extract entities: person, address, narcotic drug, and personal property. Feldman and others [16] proposed the TEG (Trainable Extraction Grammar) system that integrated a statistical and knowledge-based approach to extract named entities and relations. Maynard and others [17] also used a rule-based approach and lexical lookup to extract entities such as person, organization, location, and date from news texts. Srilahi and others [18] used machine learning with a FST (finite state transducer) rule-based approach and lexical lookup to carry out IE for question answering.

3. COGNITIVE INTERVIEW

Accurate recall of a lot of information is difficult to achieve. Most investigators are trained to rely on ‘who, what, where, when, and why’ questions [19] when they interview people. Unfortunately, in many cases this results in only a subset of the relevant information being gathered. It has been shown that a technique such as the cognitive interview (CI) [19] can help people remember more information with high accuracy.

The cognitive interview technique is based on psychological principles of memory storage and retrieval of information [20]. The first step of the cognitive interview includes mental reinstatement, a way to help witnesses mentally reconstruct the context of an incident, and encouraging witnesses to recall as much detailed information as possible. Next, interviewers ask witnesses to recall the event in different temporal orders. Finally, witnesses are encouraged to remember an incident from different perspectives, i.e., from different physical locations [20].

The cognitive interview has many advantages. Unfortunately, there are also several disadvantages. For example, it is time and labor intensive. Detectives have to be trained in advance to use this technique and it takes a long time to conduct a thorough interview.

An online interviewing system may have the same benefits as the cognitive interview while avoiding some of the problems. The system can interview people without time constraint and people can use this system at any place where computers and Internet are available.

4. SYSTEM DEVELOPMENT

Our information extraction system combines a large crime-specific lexicon, several GATE (General Architecture for Text Engineering) modules, and an algorithm to recognize the relevant entities among the phrases generated by the system.

LEXICON DEVELOPMENT

Our lexicon contains several subsets that help recognize entities, such as weapons, vehicles, scenes, clothes, shoes, and physical features. To build this lexicon we used several different resources.

We collected crime types and crime definitions from the Uniform Crime Reports (UCR). To build the vehicle and weapon lexicons, we used encyclopedia data sources such as Wikipedia1 and MSN Encarta2. To collect abstract lexicons such as scene and physical features, we used FrameNet3. To build specific lexicons such as brands of cars, web sites such as Serious Wheels4 were used. To further expand and complete our lexicons, we used thesauri dictionaries such as Collins Cobuild5, MSN Encarta, Thesaurus.com6, and Microsoft Word.


We ensured that there were no overlapping terms between the categories. We selected the most frequently used definition based on Collins Cobuild dictionary to retain or remove a term when overlapping terms occurred.

GATE MODULES

We used GATE and leveraged several of its modules and plug-ins. We adopted, without adjustment, the tokenizer, sentence splitter, part-of-speech (POS) tagger, noun chunks, and ortho-matcher. We developed our own JAPE rules and

---

2. MSN Encarta, http://encarta.msn.com/
Two teens have been reported in the armed robbery of a DeBary convenience store last month.

Our lexicons are used as gazetteer lists. We have divided our lexicon into 126 gazetteers. Each rule only uses related gazetteers rather than the entire gazetteer.

JAPE Rule—JAPE (Java Annotations Pattern Engine) [21] is a GATE-specific format to define regular expressions over annotations needed for pattern matching. We created 14 JAPE files. An example rule for the sub-lexicon ‘harm’ is:

Rule: harm

{Lookup.minorType == harm}

:harm -->

:harm.Rule = {majorType = “act”, minorType = "harm", Rule = “harm Rule”}
The section `{Lookup.minorType == harm}` matches all words from crime-related ‘harm’ acts which were stored in our gazetteer lists. So this rules indicates that if a word earlier tagged as ‘harm’ is found, it needs to be annotated as an ‘Act’ of the subtype ‘harm’.

**CRIME IE ALGORITHM**

During initial testing, we found that noun phrases sometimes provide irrelevant information such as ‘need’, ‘favor’, ‘solution’, ‘efforts’, ‘terms’, and ‘experience’. In this case, we compared the last token of the noun phrase with the gazetteer lists to filter those irrelevant noun phrases. Furthermore, the JAPE rules and the noun phrase chunker may generate overlapping phrases because they are used in parallel. In this case, we would prefer to retain only the longest phrase. For example, the phrase ‘DeBary convenience store’ contains more information than the word ‘store’ alone. Richer information can help detectives solve crimes more efficiently.

To resolve these problems, we developed a filtering algorithm (see Figure 2). First, the algorithm removes determiners such as ‘a’, ‘an’, and ‘the’ from all phrases. Next, it captures those phrases generated by the noun phrase chunker but not by the JAPE rules and obtains the last token of each phrase. If this token is not found in any gazetteer list, the algorithm discards those noun phrases since they are most probably irrelevant. We can make this assumption because the gazetteer lists enumerate all relevant entities we are interested in. Next, phrases generated by both the JAPE rules and the noun phrase chunker are evaluated. The algorithm selects the longer phrases if the phrases have different lengths. If two phrases are equal, the algorithm will discard one of them. For our example, these rules result in:

This output result is used to match questions that are pre-stored in the database. The matching is done based on the cognitive interview principles. The additional questions will lead to more information from users that will be processed in the same manner. Each time the extra information is stored in the database and additional questions are asked when necessary. We are currently testing the question generation component and report here on the information component.

**5. EVALUATION**

**METHODOLOGY**

To develop and fine-tune our system, we collected a diverse corpus containing texts from Unsolved-Crimes7, SUBA District Unit8, True Crime Blog9, Baltimore Crime10, Chat LawInfo11, and ExpertLaw12. For final testing of the information extraction modules, we collected two types of representative texts: police and witness reports. The police narratives were texts collected from alt.True-Crime13, Secret Witness14, SFGate Crime15, Crime-Stopper16, FreeAdvice17, TheLAW.com18, and LaborLawTalk19.

---

8 SUBA, http://groups.google.com/group/SPD_SDU/web/crime-bulletins
13 alt.True-Crime, http://groups.google.com/group/alt.true-crime/topics
17 FreeAdvice, http://forum.freeadvice.com/
The average length of the 20 police narratives is 130 words while that of the 20 witness narratives is 240 words. The open-source spell checker Ekit was used to correct typos in the witness narratives. The first author selected the best alternative word for each typo.

Precision was very high for both types of narratives: 94% for police narratives and 96% for witness narratives. Recall was also very high: 85% for police narratives and 90% for witness narratives. Table 1 and Table 2 provide an overview of precision and recall for each entity we extracted. For police narratives, we achieved 100% precision for ‘Personal Property’, ‘Physical Condition’, ‘Vehicle’, and ‘Clothes’, but lower precision (80%) for ‘Face’. For witness narratives, we achieved 100% precision for ‘Body Part’, ‘Physical Condition’, ‘Vehicle’, ‘Weapon’, and ‘Drug’ and encountered the lowest precision (58%) for ‘Age’. Recall for ‘Age’ was low for both police narratives and witness narratives. This is due to text such as ‘Steven Warrichiet, 40,’ and ‘14yr-Girl/15yrs-Boy’ from which the system did not extract age correctly.

The witness narratives often contain slang or street language, such as ‘weed roach’ or ‘daddy’ and unorganized syntax such as sentence fragments. Therefore we expected lower recall and precision for the witness narratives. Surprisingly, precision and recall were higher for the witness narratives than for the police narratives. A partial explanation is that, the spell checker removed most of the typos. Only one typo was found in the 20 police narratives while 51 typos were

| People          | 242 | 4  | 32  | 98% | 87% |
| Act             | 75  | 1  | 9   | 99% | 88% |
| Scene           | 101 | 20 | 5   | 83% | 80% |
| Time            | 58  | 5  | 3   | 92% | 88% |
| Age             | 7   | 1  | 12  | 88% | 35% |
| Face            | 8   | 2  | 0   | 80% | 80% |
| Body Part       | 19  | 0  | 2   | 100% | 90% |
| Personal Property | 9  | 0  | 5   | 100% | 64% |
| Physical Condition | 10  | 1  | 3   | 91% | 71% |
| Vehicle         | 29  | 0  | 0   | 100% | 100%
| Clothes         | 18  | 0  | 0   | 100% | 100%
| Weapon          | 9   | 0  | 2   | 100% | 82%
| Feature         | 10  | 1  | 3   | 91% | 71% |
| Drug            | 2   | 0  | 0   | 100% | 100%
| Hair            | 3   | 0  | 0   | 100% | 100%
| Total           | 595 | 35 | 73  | 94% | 85% |

| People          | 639 | 11 | 26  | 98% | 95% |
| Act             | 71  | 1  | 13  | 99% | 84% |
| Scene           | 99  | 12 | 7   | 89% | 84% |
| Time            | 62  | 8  | 2   | 89% | 86% |
| Age             | 7   | 5  | 1   | 58% | 54% |
| Face            | 1   | 1  | 0   | 50% | 50% |
| Body Part       | 10  | 0  | 1   | 100% | 91%
| Personal Property | 21  | 2  | 7   | 91% | 70%
| Physical Condition | 10  | 0  | 3   | 100% | 77%
| Vehicle         | 24  | 0  | 1   | 100% | 96%
| Clothes         | 1   | 0  | 0   | 100% | 100%
| Weapon          | 7   | 0  | 0   | 100% | 100%
| Feature         | 4   | 1  | 1   | 80% | 67%
| Drug            | 9   | 0  | 4   | 100% | 69%
| Hair            | 0   | 0  | 0   | NA | NA
| Total           | 965 | 41 | 66  | 96% | 90% |

6. RESULTS AND DISCUSSION
found in the 20 witness narratives without the spell checker. Additionally, many more correct items were available and extracted for ‘People’ from the witness narratives (639 items) than from the police narratives (242 items). The system extracted most pronouns that appeared in the witness narratives but could not extract some personal names such as ‘Keisharra Abercrombie’ in the police narratives.

7. CONCLUSION

We achieved high precision and recall when testing our modules with police and witness narratives. We plan to collect additional witness narratives using crime video system to further test our system and test the question-interaction components. Our final goal is to provide a reliable online crime reporting system people can use to report crime anonymously, that will encourage people to recall more crime information, and will provide a meaningful summary and a graphical result for police investigators to solve crimes more quickly and efficiently.

8. REFERENCES


