1-1-2007

Natural Language Processing and E-Government: Extracting Reusable Crime Report Information

Gondy A. Leroy
Claremont Graduate University

Alicia Iriberri '06
Claremont Graduate University

Recommended Citation

A. Iriberri and G. Leroy "Natural Language Processing and e-Government: Extracting Reusable Crime Report Information", IEEE International Conference on Information Reuse and Integration (IEEE IRI-2007), August 13-15, Las Vegas, Nevada, USA.

Natural Language Processing and e-Government: Extracting Reusable Crime Report Information

Alicia Iriberri, Gondy Leroy
School of Information Systems and Technology
Claremont Graduate University
alicia.iriberri@cgu.edu

Abstract

Crime reporting needs to be possible 24/7. Although 911 and tip-lines are the most publicized reporting mechanisms, several other options exist, ranging from in-person reporting to online submissions. Internet-based crime reporting systems allow victims and witnesses of crime to report incidents to police 24/7 from any location. However, these existing e-mail and text-based systems provide little support for witnesses' memory recall leading to reports with less information and lower accuracy. These systems also do not facilitate reuse and integration of the reported information with other information systems. We are developing an anonymous Online Crime Reporting System that is designed to extract relevant crime information from witness' narratives and to ask additional questions based on that information. We leverage natural language processing and investigative interviewing techniques to support memory recall and map the information directly to a database to support information reuse. We report on the evaluation of the Suspect Description Module (SDM) of the system. Our interface captures 70% (recall) of information from witness narratives with 100% precision. Additional modules will follow the design and development methods used with this module.

1. Introduction

Every year millions of crimes are committed in the United States. In 2003, the Federal Bureau of Investigation reported that 10.3 million property crimes and 1.38 million violent crimes were committed [1]. However, the Department of Justice reported that in the same year only half of all violent crimes and a third of all property crimes were reported [2]. Reporting crime to police is important for authorities and citizens because more accurate information allows policy makers, law enforcement officials, and police departments to control violence and allocate resources (i.e., policies, budgets, legislation, and program evaluation) more effectively. Accurate information benefits citizens as well because with it they can identify locations with high and low crime rates, take preventive measures, and make informed decisions on where to live.

Although reporting crime has many societal and individual benefits, it is common for criminal acts to remain unreported. Victims and witnesses have many reasons for not reporting a crime [3, 4]. Among these reasons, fear of repercussion, embarrassment or shame, believing the crime is too insignificant or a personal issue, believing that reporting will not make a difference, and being unable to reach an authority are often cited [5].

Researchers have proposed alternative methods to in-person reporting to increase the number of reports. Garcia and Henderson [6] describe a Blind Reporting System in which victims file reports in complete confidentiality; however, they have to be physically present at the police station to file the report. Blind reporting works under the assumption that victims are willing to report, invest the time, and bear the inconvenience of reaching authorities. Kidd and Chayet [3] indicate that victims weigh the cost and benefits of reporting criminal acts before making a decision to report. Therefore, reporting methods that reduce cost (e.g., time spent at police station) and concerns (e.g., fear of repercussions) have to be available. It is necessary to investigate alternative ways for people to report crimes and design mechanisms to ensure accessibility, anonymity, and safety of crime victims, without compromising the accuracy and completeness of crime information. Iriberri, Leroy, and Garret [7] found that people would be more willing to report crimes if they had convenient and confidential alternatives to do it, such as via the Internet.

A few initiatives to use the Internet to report crime exist. The FBI Tips and Public Leads System [8] and the Claremont University Consortium's Silent Witness Program [9] are two representative examples. Using the Internet, these systems address the concerns of victims facing the decision to report a crime. The Internet provides the convenience to reach authorities 24/7 from any location with Internet access while protecting the

victim's identity. Both systems allow victims and witnesses to report incidents using either a text-box or fill in-the-blanks input fields. They both require that the person filing the report remembers all vital information related to the crime, without support for memory recall. This lack of support results in the omission of vital information from witnesses. These e-mail and text-based systems make it difficult for police to reuse the reported information since they have to manually format it into their standard crime report form if they are to use it for further investigation.

We are developing a Crime Reporting System that will address not only the cited reasons for not reporting, but also the need of police departments for more accurate, complete, and reusable information that may free up their time and resources to allocate them to policing the streets. Our system will incorporate the convenience of the Internet, the support of techniques for memory recall, the information extraction capabilities of natural language processing (NLP) technologies and the utilities for storage, integration and reuse of information of database and electronic technologies.

With our approach, witnesses will provide more accurate and complete crime information that can be stored in a database so police can create up-to-date, ad-hoc reports and overviews, and can combine this information with other sources, such as geographic information, to provide more in-depth new insights on crime. With more information gathered, police departments will be one step closer to achieving a more effective allocation of resources to better prevent and solve crimes.

With natural language processing and the memory enhancing techniques used in investigative interviews we aim to design a Crime Reporting System that is a convenient and safe way for victims and witnesses to provide more information correctly in a format that is immediately reusable. We report here on one module of our system: the Suspect Description Module (SDM). Additional modules including location, vehicle and weapon descriptions will follow the same design and development methods as this one. We determine how effective the SDM is at processing natural language input from witnesses and matching this information to the standard police format. In particular, we measure recall and precision when extracting information about suspect descriptions from witnesses' written crime narratives.

2. Cognitive and Computer Interviews

Our system is based on the principles used in the Cognitive Interview (CI). CI is a form of investigative interview that facilitates witness memory recall. CI increases the amount of correct information obtained form witnesses anywhere from 75 to 95% compared to the information obtained using standard police interviews [10-12].

Geiselman and Fischer [13], creators of the CI, found that investigative interviews are more effective in obtaining more information when the interviewer uses triggers to facilitate witness memory recall, as opposed to only using closed and direct questioning. These triggers include letting witnesses report in narrative form every detail they can remember and helping them reinstate the context of the event they are reporting. The interviewer then uses the witness narrative to generate a strategy to ask new probing questions to complement the initial narrative, as opposed to solely asking the next question in turn from a standard police questionnaire.

Computer interviews research shows that people questioned about sensitive or embarrassing situations, such as medical history, drug or alcohol use, and sexual activity feel more comfortable answering on a computer compared to answering to a human interviewer [14-18]. Interviewees are more relaxed, open and honest with a computer system. This effect increases when the format of the questions have a human-like style [19]. Additionally, self-administered sensitive questions have shown to increases reporting levels.

Since asking to report in narrative form in interaction with computer-generated-human-style questioning has such advantages, our goal is to make it an essential part of our online Crime Reporting System. To achieve this goal we need to be able to extract information from those narratives, analyze it and use it to generate new probing questions to further the interview and obtain more correct information, just as a human CI interviewer does, and finally, store it directly in the police preferred format.

3. Suspect Description Module Development

The SDM, see Figure 1, will be the model for all modules in our system. This module prompts witnesses asking them for a description of the facial features of the crime suspect. The SDM uses NLP to analyze information in the witness narrative and extract that which is required in a standard police report. We use a standard police report as a target to ensure that information can be used without further intervention. After extracting information from the initial narrative, the SDM matches the extracted information with the standard report to identify missing suspect characteristics, and generates further probing questions to prompt the witness asking them for the remaining suspect characteristics.

The SDM was developed using a Java-based user interface and the information extraction tools from the General Architecture for Text Engineering (GATE) system[20].
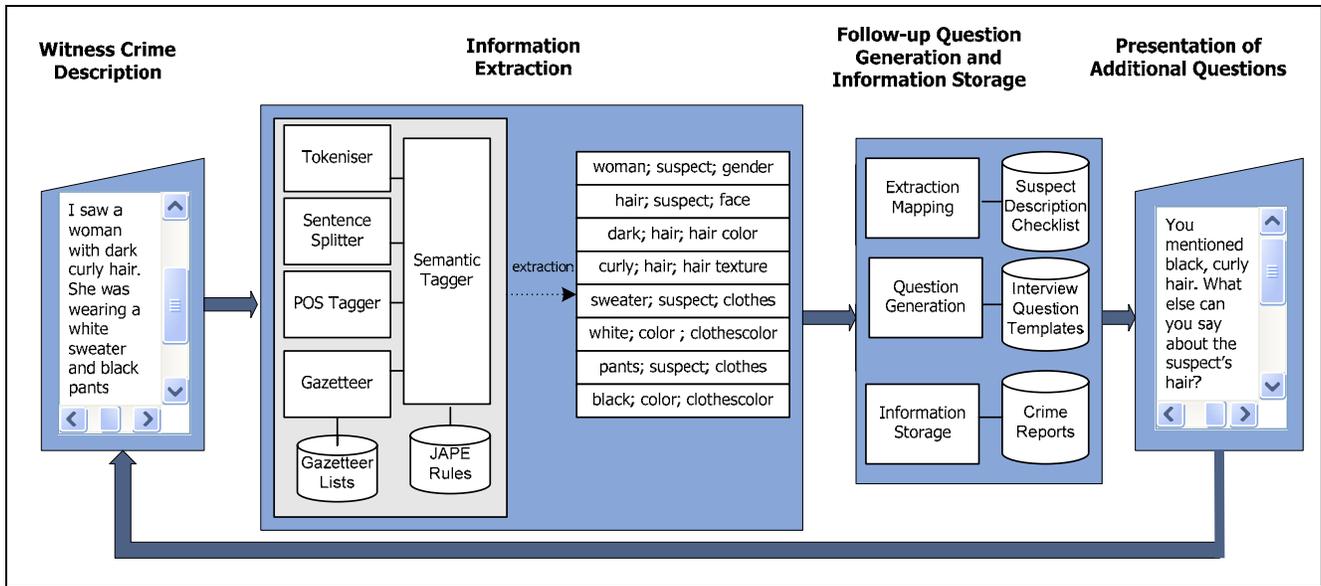
Figure 1. Suspect Description Module Overview

GATE is a readily available open source system that facilitates the creation of information extraction applications. Various approaches to Information Extraction exist, namely lexical lookup, rule-based, statistical-based and machine learning. We chose a combination of the first two approaches given that we had limited access to crime reports to use as training cases. Furthermore, GATE is a generic and convenient resource to use in the prototyping stages of our Crime Reporting System.

The SDM interface presents a text box and asks the witness to provide as much information as she can recall about the suspect facial features in the same way a cognitive interviewer would do it. The narrative is stored as a text file and becomes the input for the subsequent information extraction tasks. The narrative is first separated into text units or tokens and sentences. The tokens are then tagged with part-of-speech annotations. These tokens are matched with predefined lexicons that contain lists of face features and their attributes, and finally sentences are analyzed using hand-crafted rules that describe common text patterns. We use the following GATE's Information Extraction tools as a basis to perform some of these tasks:

- **Tokeniser**. This resource splits the text in the witness narrative into tokens such as numbers, punctuation, and words.
- **Sentence Splitter**. The sentence splitter segments the narrative text into sentences. This is a necessary step for future part-of-speech identification.
- **Part of Speech (POS) Tagger**. The tagger annotates tokens with part-of-speech tags, for example, nouns and adjectives. These tags help to identify characteristics such as hair and hair color in the narrative text.

- **Gazetteer**. This resource facilitates the identification of entities (e.g., facial features) in the input text (i.e., narratives). Tokens in the input text are compared with entries in Gazetteer lists to annotate, for example, face parts, hair color, and hair texture. We created 39 Gazetteers with facial parts and their attributes. Table 1 lists examples of the contents of the Gazetteer list for hair color. The Gazetteer tables were populated using several sources. One was a sample of standard suspect-description formats used by actual police departments. Other sources were WordNet [21] and various Internet sites that list people's facial features such as eye colors, skin complexions and nose shapes.

Table 1. SDM's Attribute's list for Hair Color

blond	brown	dark brown
dark blond	dark	light brown
light blond	white	gray
black	red	sandy brown

- **Semantic Tagger**. This resource uses sets of rules or Java Annotation Pattern (JAPE) rules, which are hand-crafted and stated as regular expressions. These rules act as finite state transducers to annotate input text assigning semantic tags that correspond to specific required information. For example, a noun phrase that includes "brown" and "hair" would be annotated as "hair color". We created 85 JAPE rules to define facial characteristics. The rules were tested and fine-tuned using a sample of 45 narratives collected during two

pilot tests. Table 2 shows an example rule (additional, example Gazetteers and JAPE rules will be available at http://isl.cgu.edu.). If in the input text a noun phrase, such as "brown hair" where the word "brown" was previously tagged as "haircolor" is encountered, the JAPE transducer will trigger this rule. The result will be the string "brown hair; hair; hair color." The SDM will then take this output string and check the feature hair color in the suspect description checklist.

Table 2. SDM's JAPE Rule for Suspect's Hair Color

```
Rule: HairColor
    ( {Lookup.minorType == haircolor}
      ({SpaceToken.kind == space})?
      {Token.string == "hair"}    )
  :HairColor -->
  :HairColor.Rule = {majorType = "hair",
minorType = "haircolor"}
```

4. Suspect Description Module Evaluation

At this stage, we are evaluating the information extraction efficiency of the SDM compared to manual human information extraction. Given a witness narrative and a standard police suspect-description form, we created a suspect description checklist which is our gold standard. The SDM is expected to extract all the features in the narrative that match the checklist elements. Information that would not map to the standardized checklist will be stored separately for inclusion in the final crime report. We evaluate here the first round of information extraction, since that will be indicative of further performance of the system. Probing questions were not posed to witness for this evaluation. These questions will be posed of missing information and subsequent narratives will be treated as the first. Questions will be pre-generated for each possible missing item.

4.1. Methodology

4.1.1.Test Set Creation. We obtained eight 8x11inch-color pictures depicting faces of actual crime suspects. The pictures were downloaded from the "Most-Wanted" website www.placer.ca.gov/Sheriff/MostWanted.aspx. We then asked friends and family to write narratives describing the individuals in the pictures. This is a similar activity to one performed in real interviews. The individuals in the pictures were of four different ethnicities and were paired by gender as follows, two African-Americans, two Hispanics, two Asians, and two Caucasian.

Thirty-one individuals, males and females ranging in age from 18 to over 50 years old and in education level from some-college to post-graduate levels provided us with narratives to use in this system evaluation. We randomly assigned one picture to each individual and asked them to fill out a brief questionnaire which included a question asking them to describe the facial features of the individual in the picture. We showed the picture to the participants for 20 seconds and asked them to answer their questionnaire without looking back at the pictures.

4.1.2.Gold Standard Mapping. Each of the 31 narratives was read to identify physical features listed, such as eye color, hair color, hair texture and facial hair. For each narrative, we manually extracted all information related to the suspect's physical descriptions. We then mapped the extracted information to the checklist and counted the number of matching features in each narrative to determine the total number of relevant (i.e., required) features in each input text. We also tagged each identified feature accordingly. The resulting lists became the gold standard for each of the narratives to use in the evaluation of the SDM. The SDM extraction would have to match these gold standards.

4.1.3.Precision and Recall Calculation. To evaluate the information extraction performance of the SDM we measured its average recall and precision. We define *recall* as the number of features in the input text that the SDM extracts out of the number of relevant features included in the text and *precision* as the number of features that are correctly identified and annotated out of the total number of features that were extracted from the input text, including features incorrectly identified or not identified at all. That is,

$$\text{recall} = \frac{\text{required and extracted features}}{\text{required features}}$$

$$\text{precision} = \frac{\text{required and extracted features}}{\text{extracted features}}$$

4.2 Results

We typed and entered individually each of the 31 narratives into the SDM for processing. Each narrative text ranged from 21 to 101 words in length with an average of 59 words. For each narrative, the SDM generated a list of relevant features annotated by suspect facial characteristic, face part and part's attributes. For instance "blond hair" would be listed as: *"blond hair";* and annotated as: *hair; hair color.* "Female Caucasian" would be listed as: *"Female"* and tagged as: *suspect; gender* and *"Caucasian"* as: *suspect; race*.

The number of extracted features from each narrative's relevant set of features and the number of correctly annotated (or correctly identified) features were counted.

Table 3. SDM Recall and Precision

Recall Level	Total Number (%) of Narratives	Avg. Extracted Features	Avg. Required Features	Recall	Precision
Less than 50%	3 (10%)	3	11	32%	100%
50 - 70%	11 (35%)	7	12	54%	100%
71 - 90%	12 (39%)	8	10	79%	100%
Greater than 90%	5 (16%)	10	11	94%	100%
Overall	**31**	**7**	**11**	**70%**	**100%**

Resulting recall and precision rates for the SDM are presented in Table 3. The table also shows the distribution of narratives to recall. The results for the automatic extraction with the SDM after pilot tests and this final evaluation were 70% recall and 100% precision.

5. Discussion

The main reason for the moderate recall level (70%) were missing attributes in our Gazetteer lists and missing JAPE rules to extract information from phrases with complex structure or complex noun-attribute combinations. Table 4 lists the number of cases and the reasons why the SDM was not able process completely.

Table 4. Reasons for Moderate Recall

Number of Narratives	Reason Category
15	Missing Attribute in Gazetteer Lists
3	Missing Phrase Pattern in JAPE Rules
10	Missing Attributes and Missing Phrase Patterns
3	No Problem - Full Extraction
31	**Total**

The omission of attributes and rules is something that can be continuously improved during the next development cycles. We can add missing attributes and synonyms in the Gazetteer tables and add JAPE rules to identify text patterns that include the complex sentence structures we encountered. Missing attributes included colors such as "orange" or "lavender" or clothes styles like "dress shirt." These were, therefore, not identified by the SDM.

The SDM was generally successful in extracting information from narratives that use simple and deterministic noun phrases. For example, phrases like "White female," "dark brown hair," or "medium length hair" were consistently and correctly extracted and identified by the SDM. On the other hand, phrases like "face somewhat red," "hair beginning to grow back in," "nose seemed red," although relevant, were not extracted. The SDM had problem identifying features that were described using complex noun phrases, for example, those that use correlative conjunctions like "her lips were not big but full", or those that describe more that one noun like "crooked mouth and nose."

In terms of precision, the SDM was able to classify clearly all the features that it was able to extract. For example, terms such as "he," "female," and "man" were classified as suspect's gender, and "35-45 years old" or "in her 30's" were always correctly classified regardless of format as the suspect's age. Those features that would not be possible to classify were not extracted in the first place by the SDM as it was not able to identify them.

Another interesting finding was the result of allowing participants to write as many details as they recalled and allowing them to write a free flowing narrative. This resulted in a richness of information that they would otherwise not provide according to research with the Cognitive Interview. For example, a participant wrote "he had blue piercing eyes" as opposed to just answering "blue eyes" as she would have done when asked "what color were the suspect's eyes?" Moreover, participants were able to include references in their descriptions such as "she looked like a school girl," or "he looked like an actor."

The data collected in this evaluation and the experience of testing the behavior of the SDM will be

used to further refine and complement the JAPE rules used for information extraction as we develop our Crime Reporting System.

6. Conclusions

We reported on the design and evaluation of one module of our Crime Reporting System: the Suspect Description Module. This module is based on Cognitive Interview techniques and Natural Language Processing. The module leverages GATE and uses 39 Gazetteers and 85 pattern rules to extract information from witnesses' narratives.

We showed pictures of crime suspects' faces to 31 individuals and asked them to provide us with written descriptions of those suspects. We processed those narratives both manually and automatically using the SDM. The results of this evaluation showed very high precision and moderate recall. These results will be used to fine-tune the SDM and to guide the design of additional modules of our Crime Reporting System.

References

1. FBI, *Crime in the USA*. 2003, Federal Bureau of Investigation.
2. BJS, *Reporting Crime to the Police, 1992-2000, Bureau of Justice Statistics Special Report*. 2005.
3. Kidd, R.F. and E.F. Chayet, *Why do victims fail to report? The psychology of criminal victimization.* Journal of Social Issues, 1984. **40**: p. 39-50.
4. Singer, S., *The Fear of Reprisal and the Failure of Victims to Report a Personal Crime.* Journal of Quantitative Criminology, 1988. **4**(3): p. 289-302.
5. Thomson, R. and J. Langley, *Who Do Young Adult Victims of Physical Assault Talk to about their Experiences?* Journal of Community Psychology, 2004. **32**(4): p. 479-488.
6. Garcia, S. and M. Henderson, *Blind Reporting of Sexual Violence.* FBI Law enforcement bulletin, 1999. **68**(6): p. 12-16.
7. Iriberri, A., G. Leroy, and N. Garret. *Reporting Crime Online: User Intention to Use*. in *39th Hawaii International Conference on Systems (CD-ROM) Science, January 4-7*. 2006. Kauai, HI: Computer Society Press.
8. FBI, *FBI Tips and Public Leads*. 2005, Federal Bureau of Investigations.
9. CUC, *CUC Silent Witness Program. Campus Safety, Claremont Graduate University*. 2005.
10. Fisher, R.P., et al., *Enhancing Enhanced Eyewitness Memory: Refining the Cognitive Interview.* Journal of Police Science and Administration, 1987. **15**(4): p. 291-297.
11. Fisher, R.P., R.E. Geiselman, and M. Amador, *Field test of the Cognitive Interview: Enhancing the Recollection of Actual Victims and Witnesses of Crime.* Journal of Applied Psychology, 1989. **74**: p. 722-727.
12. Geiselman, R.E., et al., *Enhancing eyewitness Memory: Refining the Cognitive Interview.* Journal of Police Science and Administration, 1987. **15**(4): p. 292.
13. Fisher, R.P. and R.E. Geiselman, *Memory-enhancing techniques for investigative interviewing: The cognitive interview*. 1992, Springfield, IL: Charles C Thomas.
14. Tourangeau, R. and T.W. Smith, *Asking Sensitive Questions: The Impact of Data Collection Mode, Question Format and Question Context.* The Public Opinion Quarterly, 1996. **60**(2): p. 275-304.
15. Bachman, J.W., *The Patient-Computer Interview: A Neglected Tool that Can Aid the Clinician.* Mayo Clinic Proceedings, 2003. **78**: p. 67-78.
16. Waterton, J.J. and J.C. Duffy, *A Comparison of Computer Interviewing Techniques and Traditional Methods in the Collection of Self-Report Alcohol Consumption Data in a Field Survey.* International Statistical Review, 1984. **52**(2): p. 173-182.
17. Lessler, J.T. and J.M. O'Reilly, *Mode of Interview and Reporting of Sensitive Issues: Design and Implementation of Audio Computer-Assisted Self-Interview.* National Institute on Drug Abuse: Research Monograph, 1997. **167**: p. 366-382.
18. Greist, J.H., et al., *Comparison of Computer and Interviewer-Administer Versions of the Diagnostic Interview Schedule.* Hosp Community Psychiatry, 1987. **38**: p. 1304-1311.
19. Peiris, R., P. Gregor, and N. Alm, *The Effects of Simulating Human Conversational Style in a Computer-based Interview.* Interacting with Computers, 2000. **12**: p. 635-650.
20. Cunningham, H., et al. *GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications*. in *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*. 2002. Philadelphia.
21. Fellbaum, C., ed. *WordNet: An Electronic Lexical Database*. 1998, The MIT Press.