Extracting Meaningful Entities from Police Narrative Reports

Michael Chau, Jennifer J. Xu, Hsinchun Chen
Department of Management Information Systems
The University of Arizona
Tucson, AZ 85721, USA
{mchau, jxu, hchen}@bpa.arizona.edu

Abstract

Valuable criminal-justice data in free texts such as police narrative reports are currently difficult to be accessed and used by intelligence investigators in crime analyses. It would be desirable to automatically identify from text reports meaningful entities, such as person names, addresses, narcotic drugs, or vehicle names to facilitate crime investigation. In this paper, we report our work on a neural network-based entity extractor, which applies named-entity extraction techniques to identify useful entities from police narrative reports. Preliminary evaluation results demonstrated that our approach is feasible and has some potential values for real-life applications. Our system achieved encouraging precision and recall rates for person names and narcotic drugs, but did not perform well for addresses and personal properties. Our future work includes conducting larger-scale evaluation studies and enhancing the system to capture human knowledge interactively.

1. Introduction

Efficient and effective access of criminal-justice data is critical for law enforcement personnel to perform investigations and fight crimes. Currently, most criminal-justice data are stored in structured relational databases, in which data are represented as tables consisting of various fields, such as the attributes of a suspect, the address of a crime scene, and so on. Detectives and crime investigators can search for useful information in such databases by providing specific search queries (e.g., Chen et al., 2002). These databases, however, may not be comprehensive because they can only capture structured data which fit in the pre-defined fields and attributes in the databases. Unstructured data, such as free-text police narrative reports, often are stored as text objects in databases or simply text files. Valuable information in such texts is difficult to be accessed or used by crime investigators in further analyses. It would be useful to identify from such text reports meaningful entities, such as person names, narcotic drugs, or personal properties. Extracting such entities, a process usually referred to as named-entity extraction, will not only help crime investigation processes but also provide necessary input for data mining applications such as crime pattern recognition and criminal relationship identification (Hauck et al., 2002).

In this paper, we report our work on an entity extractor, which utilizes named-entity extraction techniques to identify meaningful entities from police narrative reports. The rest of the paper is organized as follows. In Section 2, we provide an overview on named-entity extraction research. We outline our research questions in Section 3. In Section 4, we discuss the system architecture adopted in our study. In Section 5, we report the design and results of a preliminary experiment conducted to evaluate the performance of our system. In the last section, we discuss the main findings of this research and some possible future research directions.

2. Background

Named-entity extraction, a subfield of information extraction, refers to the automatically identification from text documents the names of entities of interest, such as persons (e.g., “John Doe”), locations (e.g., “Washington, D.C.”), and organizations (e.g., “National Science Foundation”). It has also been extended to identify other patterns, such as dates, times, number expressions, dollar amounts, email addresses, and

Web addresses (URLs). The Message Understanding Conference (MUC) series has been the major forum for researchers in this area, where they meet and compare the performance of their entity extraction approaches (Chinchor, 1998).

There are four major named-entity extraction approaches: lexical-lookup, rule-based, statistic-based, and machine learning.

- *Lexical lookup*: Most entity extraction systems maintain hand-crafted lexicons that contain lists of popular names for entities of interest, such as all registered organization names in the U.S., all person last names obtained from the government census data, etc. These systems work by looking up phrases in texts that match the items specified in their lexicons (e.g., Borthwick et al., 1997).
- *Rule-based*: Rule-based systems rely on hand-crafted rules to identify named entities. The rules may be structural, contextual, or lexical (Krupka & Hausman, 1998). An example rule would look like the following:

  $$\textit{capitalized last name} + \text{, } + \textit{capitalized first name} \Rightarrow \textit{person name}$$

  Although such human created rules are usually of high quality, this approach is not scalable to large number of entities and not robust to less structured documents.
- *Statistic-based*: Systems in this category often use statistical models to identify occurrences of certain cues of particular patterns for entities in texts. A training data set is needed for these systems to obtain the statistics. The statistical language model (Witten et al., 1999) is an example of statistic-based entity extraction systems.
- *Machine learning*: This type of systems relies on machine learning algorithms rather than human created rules to extract knowledge or identify patterns from texts. Examples of machine learning algorithms include neural networks, decision tree (Baluja et al., 1999), Hidden Markov Model (Miller et al., 1998), and entropy maximization (Borthwick et al., 1998).

Instead of relying on a single approach, most existing information extraction systems utilize a combination of two or more of these approaches. Many systems that used a combined approach were evaluated at the MUC-7 conference. The best systems were able to achieve over 90% in both precision and recall rates in extracting persons, locations, organizations, dates, times, currencies, and percentages from a collection of New York Times News (Chinchor, 1998).

3. Research Questions

As discussed earlier, police narrative reports contain a substantial amount of useful information that may not be available in structured databases. However, such information is usually buried in unstructured text documents, thus limiting its usage in crime investigation and analysis. It would be desirable to apply information extraction techniques such as named-entity extraction to such reports. Despite the availability of generic entity extraction systems, analysis of narrative reports poses two new challenges. First, narrative reports are very noisy compared to other types of text data such as news articles. They usually contain many spelling errors, grammatical mistakes, and typos, thus making the entity extraction task more difficult. Second, entities other than person names, organizations, and locations, such as drug names, crime addresses, and vehicles, are also equally relevant to crime intelligence analysis. In this research, we report our work to address these problems such that various useful entities can be identified from relatively noisy narrative reports by applying entity extraction techniques.

4. System Architecture

Our entity extractor is a neural network-based system. Rather than relying entirely on manual rule generation, this system combines lexical lookup, machine learning, and minimal hand-crafted rules. It has three major components:

- *Noun phrasing*: This component is a modified version of the Arizona Noun Phraser (Tolle & Chen, 2000). It extracts noun phrases from documents based on syntactical analysis. The noun phrases extracted are candidates for named entities.
- *Finite state machine and lexical lookup*: A finite state machine processes the noun phrases identified by the Arizona Noun Phraser. It compares each word in the phrase, as well as the word immediately before and the word immediately after the phrase, with the items in the hand-crafted lexicons. Each comparison will generate a binary value (either 0 or 1) to indicate a match or mismatch. Some other scores, such as a word feature score calculated based on uppercase/lowercase information, are also generated for later use.
- *Neural network*: The feedforward/backpropagation neural network component takes a phrase's corresponding binary values and the scores generated by the finite state machine as input, and then predicts the phrase's most possible entity type.

The system has two states: a training state and a testing state. In the training state, the system identifies lexical rule patterns based on a training data set, which consists of manually tagged input-output pairs. These learned patterns are stored as synaptic weights in the neural network (Lippmann, 1987). In the testing state, the system extracts phrases from testing documents and predicts the entity type of each phrase based on the lexical rule patterns obtained from the learning state.

Originally, our entity extractor was designed to recognize only three entity types: person, location, and organization. After consulting with several detectives and crime analysts at the Tucson Police Department, we identified five entity types that the crime investigators believed would be useful in their crime investigations. The five types are person, address, vehicle, narcotic drug, and personal property. In order to customize the system for these types, we created the lexicons and designed the finite state machines for the four new entity types. The lexicons were created mainly based on the structured database records at the Tucson Police Department.

5. Evaluation

5.1 Testbed

We designed and conducted an experiment to evaluate the usefulness and performance of our entity extractor. A set of police narrative reports from the Phoenix Police Department was used as the testbed in our preliminary evaluation. The testbed consisted of 36 reports randomly selected from the Phoenix Police Department database for narcotic related crimes. A human experimenter manually tagged these reports to identify all entities that belong to the five entities of interest. The narrative reports were relatively noisy: they all were written in uppercase letters (thus lacking the letter case information, which appears to be very useful in named-entity extraction for English documents). The reports also contained a significant amount of typos, spelling errors, and grammatical mistakes. This added difficulty to the entity extraction process. On the other hand, however, it allowed us to test our system's robustness for noisy data sets.

5.2 Preliminary Results

We performed a 3-fold cross validation testing on the system, a common method used in evaluation of machine learning applications. In our 3-fold cross validation, the 36 reports were divided into three sets and the validation was conducted in three iterations. In the first iteration, the first and the second sets were used for training, while the last set for testing. In the second iteration, the first and last sets were used for training, and the second for testing. In the last iteration, the second and last sets were used for training and the first for testing.

We used precision and recall rates to measure the performance of our system. They were calculated as follows:

$$precision = \frac{number\ of\ correct\ entities\ identified\ by\ the\ system}{number\ of\ all\ entities\ identified\ by\ the\ system} \quad (1)$$

$$recall = \frac{number\ of\ correct\ entities\ identified\ by\ the\ system}{number\ of\ entities\ identified\ by\ human} \quad (2)$$

The results are summarized in table 1.

	Precision	Recall	Number of correct entities extracted by system	Number of total entities extracted by system	Number of total entities extracted by human
Person	0.741	0.734	429	617	600
Address	0.596	0.514	30	52	62
Narcotic drug	0.854	0.779	200	233	252
Personal property	0.468	0.478	137	350	291

*The result for "Vehicle" was not included because there were only four occurrences of vehicles in the 36 reports.

Table 1. Experimental results

In general, the results indicate that our entity extractor performs well in identifying person names and narcotic drugs from the test data set. Although the precision and recall rates for these two types were not as good as the results reported at MUC, we were satisfied considering that the narrative reports were much noisier than the news articles used in MUC evaluation. However, the entity extractor's performance for addresses was not as good as we expected. After examining the test data, we found that the unsatisfactory performance for address identification was due to mistakes in the manually created lexicons. We believed that the performance could be sharply improved if the errors in the lexicons were fixed. On the other hand, the poor performance for personal property identification was somewhat expected. Personal properties were much harder to identify because they may include any physical object that a person can own, such as TVs, social security cards, car battery, sofa, and so on.

6. Conclusion and Future Directions

In this paper, we report our work on applying named-entity extraction techniques to identifying useful entities from police narrative reports. Preliminary evaluation results have demonstrated the feasibility and the potential values of our approach. In the future, we plan to use larger training and testing data sets to evaluate the performance of the system. We are also researching into how we can use an interactive approach to integrate human knowledge into the system's learning component.

Acknowledgement

We would like to thank the National Science Foundation (NSF) and the National Institute for Justice (NIJ) for funding this project. We would also like to thank the Phoenix Police Department for providing us with the narrative data. Special appreciation goes to Kristin Tolle for her significant contribution in developing an early version of the system when she was studying at and funded by the University of Arizona. We also thank Jing Zhang and Dr. Homa Atabakhsh from the University of Arizona, and Lieutenant Jennifer Schroeder and Detective Tim Petersen from the Tucson Police Department, for their participation in this project.

References

S. Baluja, V. Mittal, and R. Sukthankar (1999). Applying machine learning for high performance named-entity extraction, in *Proceedings of the Conference of the Pacific Association for Computational Linguistics*, 1999.

A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman (1998). NYU: Description of the MENE named entity system as used in MUC-7, in *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, April 1998.

H. Chen, J. Schroeder, R. V. Hauck, L. Ridgeway, H. Atabakhsh, H. Gupta, C. Boarman, K. Rasmussen, and A. W. Clements (2002). COPLINK Connect: Information and knowledge management for law enforcement, *Decision Support Systems*, Special Issue on Digital Government, forthcoming.

N. A. Chinchor (1998). Overview of MUC-7/MET-2, in *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, April 1998.

R. V. Hauck, H. Atabakhsh, P. Ongvasith, H. Gupta, and H. Chen (2002). Using Coplink to analyze criminal justice data, *IEEE Computer*, 35(3), pp. 30-37.

G. R. Krupka and K. Hausman (1998). IsoQuest Inc.: Description of the NetOwl™ extractor system as used for MUC-7, in *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, April 1998.

R. P. Lippmann (1987). An introduction to computing with neural networks, *IEEE Acoustics Speech and Signal Processing Magazine*, 4(2), pp. 4-22.

E. Marsh and D. Perzanowski (1998). MUC-7 evaluation of IE technology: Overview of results, in *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, April 1998.

S. Miller, M. Crystal, H. Fox, L. Ramshaw, R. Schwartz, R. Stone, R. Weischedel, and the Annotation Group (1998). BBN: Description of the SIFT system as used for MUC-7, in *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, April 1998.

K. M. Tolle and H. Chen (2000). Comparing noun phrasing techniques for use with medical digital library tools, *Journal of the American Society for Information Science*, 51(4), pp. 352-370.

I. H. Witten, Z. Bray, M. Mahoui, and W. J. Teahan (1999). Using language models for generic entity extraction, in *Proceedings of the ICML Workshop on Text Mining*, 1999.