Extracting Structured Information via Automatic + Human Computation

Ellie Pavlick and **Chris Callison-Burch**
University of Pennsylvania

Abstract

We present a system for extracting structured information from unstructured text using a combination of information retrieval, natural language processing, machine learning, and crowdsourcing. We test our pipeline by building a structured database of gun violence incidents in the United States. The results of our pilot study demonstrate that the proposed methodology is a viable way of collecting large-scale, up-to-date data for public health, public policy, and social science research.

Motivation

The majority of information is encoded in the form of natural language. But structured formats, like tables or relational databases, make it easier to perform quantitative analyses of data and draw evidence-based conclusions. For many social science disciplines, the type structured information necessary for such analyses is rarely available at the scale needed. However, massive amounts of new data are made available every day in the form of unstructured language, for example in online newspapers, blogs, and public government records. Extracting structured information from these sources could enable policy makers and scientists to answer more questions rigorously and empirically.

Current state-of-the-art systems for fully-automatic information extraction (Berant et al. 2013) perform well below the level required to extract the accurate, detailed information needed for proper scientific research. We propose a hybrid methodology which combines automatic techniques (including machine learning, natural language processing, and information extraction) with crowdsourcing in order to extract detailed, structured data from natural language text. We apply our methodology to a case study in which we build a database of gun violence incidents across the US from local news articles.

Gun Violence Use Case

Gun violence provides a especially poignant example of an area of research that can benefit from our proposed methodology. Guns account for ≈33,000 deaths in the US every year, but there is no single database that enumerates the details of gun violence incidents (FICAP 2006). Although a large-scale gun violence database could enable data-driven reasoning about a topic that is usually dominated by emotion, research in this area is massively underfunded and actively blocked by federal legislation (Kassirer 1995). However, local newspapers and television stations report on gun injuries and fatalities. The details of these reports would be valuable to epidemiologists if they were in a structured database, rather than spread across the text of thousands of web pages.

This work describes a general methodology which can be applied to gun violence, or any other events for which detailed global data is difficult to access, but which is described in a dispersed fashion on the web.

Proposed Methodology

Our database population pipeline consists of 6 steps, 4 which are performed automatically and 2 which are performed by humans. Figure 1 depicts this pipeline schematically. The stages are as follows:

1. Automatic: Perform a daily web crawl of more than 2,500 local newspapers[1] covering all 50 states.
2. Automatic: Predict whether or not each of the articles in our web crawl describes an incident of gun violence using a statistical text classifier.
3. Crowd: Recruit crowd-workers to validate the predictions of our classifier, to ensure that the collected gun violence articles are high-precision.
4. Automatic: Run a suite of natural language processing (NLP) tools such as named entity recognizers and key word extractors over the validated texts. These automatic methods can facilitate the job of the crowd-workers in the next step.
5. Crowd: Recruit crowd-workers to read the articles and answer questions to populate the database. For the gun violence database, fields include information like date and location of the incident, type of weapon, and descriptions of the shooters and victims.
6. Automatic: Perform a heuristic de-duplication step in order to filter out or merge information from redundant articles.

We run a pilot study of the KBP component of our pipeline by having workers extract information from 8,800

[1] http://newspapermap.com

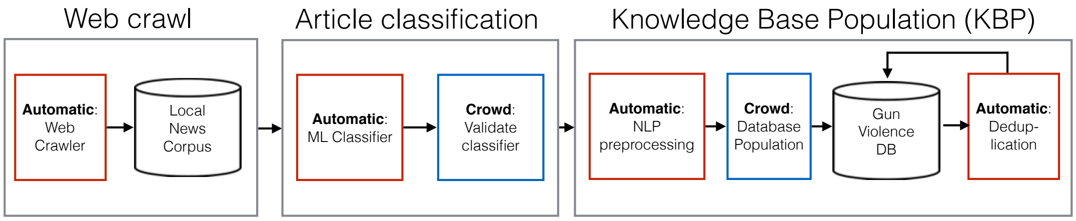
Web crawl
Article classification
Knowledge Base Population (KBP)
Automatic: Web Crawler
Local News Corpus
Automatic: ML Classifier
Crowd: Validate classifier
Automatic: NLP preprocessing
Crowd: Database Population
Gun Violence DB
Automatic: Dedup-lication

Figure 1: Pipeline for extracting structured information from text using both automatic processing (red) and human computation (blue).

articles describing gun violence scraped from the Gun Report blog[2]. In total, 505 workers processed all of our articles, taking an average of just 5 minutes per article.

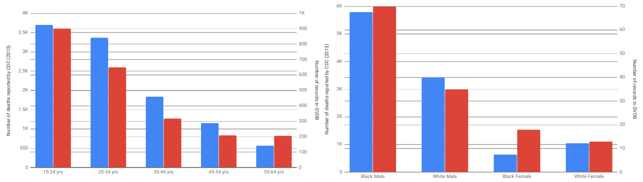

Figure 2: Deaths reported by (CDC 2013) (blue) and records in the Gun Violence Database (red), by victim's age (left) and race (right).

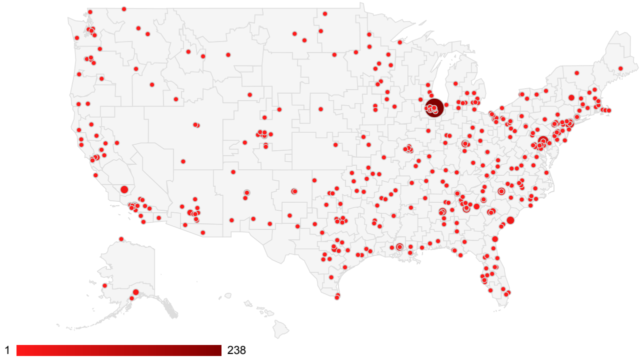

Figure 3: Reports of shootings across US cities, according to 8,800 records in our constructed database.

Results and Discussion

The database we construct from our 8,800 articles covers all 50 states and over 3,000 cities. Figure 3 shows how the records in the database are distributed across the country. By extracting data from local news reports, we are able to collect the type of locally-aggregated information which is of particular interest to social science and policy researchers. The extracted data covers nearly 2,000 reports of unintentional shootings, 565 reports of domestic violence, and over 300 reports of police-related shootings. This type of fine-grained information is especially relevant for making informed policy decisions, but is nearly impossible to extract automatically using the current state-of-the-art NLP technologies. Figure 2 compares the demographic breakdown of the homicide reports in our database to national statistics released by the CDC. The figures show that the extracted data provides good coverage of many age, race, and gender groups relevant to experts at the CDC.

Related Work

We build on a large body of work on citizen science, particularly for improving social policy (López Moncada, Farzan, and Clift 2014). Especially relevant to our work are efforts which have invoked human-in-the-loop algorithms (Hernandez et al. 2014). There are several related efforts which recruit volunteers to collect information about gun violence and police brutality (Burghart 2014; Wagner 2014).

References

Berant, J.; Chou, A.; Frostig, R.; and Liang, P. 2013. Semantic parsing on freebase from question-answer pairs. In *EMNLP*, 1533–1544.

Burghart, B. D. 2014. What Ive learned from two years collecting data on police killings.

CDC. 2013. Deaths: Final data for 2013. *National vital statistics reports: from the Centers for Disease Control and Prevention, National Center for Health Statistics, National Vital Statistics System* 64(2).

FICAP. 2006. *Firearm injury in the US.* Online Resource Book from The Firearm and Injury Center at Penn. www.uphs.upenn.edu/ficap/resourcebook/pdf/monograph.

Hernandez, A. M.; Hochheiser, H. S.; Horn, J. R.; Crowley, R. S.; and Boyce, R. D. 2014. Testing pre-annotation to help non-experts identify drug-drug interactions mentioned in drug product labeling. In *HCOMP Citizen+X Workshop*.

Kassirer, J. P. 1995. A partisan assault on science–the threat to the cdc. *New England journal of medicine* 333(12):793–794.

López Moncada, C. A.; Farzan, R.; and Clift, S. 2014. Measuring impact of local community initiatives: A crowdsourcing approach. In *HCOMP Citizen+X Workshop*.

Wagner, K. 2014. We're compiling every police-involved shooting in America. Help us.

[2] `http://nocera.blogs.nytimes.com/category/gun-report/`