

Information Extraction Approaches to Unconventional Data Sources for “Injury Surveillance System”: the Case of Newspapers Clippings

Berchiolla Paola, Scarinzi Cecilia, Snidero Silvia, Rahim Yousif, Gregori Dario



Working paper: 02/2007

THIS PAPER CAN BE CITED AND DISTRIBUTED, UNDER THE CONDITION THAT: (I) THE SOURCE IS FULLY RECOGNIZED, (II) NO MODIFICATION ARE MADE TO THIS PDF, AND (III) ANY CITATION TO THIS PAPER MUST BE STRUCTURED AS FOLLOWS:
Authors, Title, I2C Research Group Working Papers, Num: Year, www.i2crg.org.

Information Extraction Approaches to Unconventional Data Sources for “Injury Surveillance System”: the Case of Newspapers Clippings

Berchiolla Paola.¹, Scarinzi Cecilia.², Snidero Silvia.³, Rahim Yousif⁴, Gregori Dario.¹,

¹Dept. of Public Health and Microbiology, University of Torino, Italy

²Dept. of Statistics and Applied Mathematics D. de Castro, University of Torino, Italy

³S&A S.r.l., Cuneo, Italy

⁴International Society for Violence and Injury Prevention, Norway

Abstract

Injury Surveillance Systems based on traditional Hospital records or clinical data have the advantage of being a well established, highly reliable source of information for making an active surveillance on specific injuries, like choking in children. However, they suffer the drawback of delays in making data available to the analysis, due to inefficiencies in data collection procedures. In this sense, the integration of clinical based registries with unconventional data sources like newspaper articles and blogs has the advantage of making the system more useful for early alerting. Usage of such sources is difficult since information is only available in the form of free natural-language documents rather than structured databases as required by traditional data mining techniques. Information Extraction (IE) addresses the problem of transforming a corpus of textual documents into a more structured database.

In this paper, on a corpora of newspapers articles related to foreign body injuries in children we compared the performance of three IE algorithms- (i) a classical rule based system which requires a manual annotation of the rules of ; (ii) a rule based system which allows for the automatic building of rules; (iii) a machine learning method based on Support Vector Machine.

Although some useful indications are extracted from the newspaper clippings, this approach is at the time being far from being routinely implemented for injury surveillance purposes.

1. Introduction

Injury Surveillance Systems are an essential part of an effective Public Health and safety strategy (Centers for Disease Control and Prevention 2001). Common sources of injury surveillance data are traditional clinical derived data. They include hospital discharge records and death certificates (Voight, Lapidus et al. 1998) and contain information regarding the nature of injury including demographic information relating to the victim and variable degree of detail regarding the cause and the circumstances surrounding the injury.

On the other hand fatal injuries, especially those involving children, are frequently reported to the public by the media. Newspapers report injuries as part of local and national sections have been suggested for use as an injury surveillance tool being an alternative source for injury data (Fine, Jones et al. 1998). Newspaper articles reporting injuries in fact may provide details about injury circumstances and causes that are not usually available (Baullinger, Quan et al. 2001; Guard and Gallagher 2005).

Information extraction is the task concerned with identifying predefined types of information stored in natural language texts such as newspaper articles or WEB pages.

The simplest IE technology is the Named Entity Recognition (NER). NER systems identify all the names specified in users-defined lists or throughout some rules. Such lists of relevant information looked for as well as creating rules must be done manually. While creating rules and list of words is difficult and time-consuming, over the past years, a number of IE techniques have been developed to address this problem.

Successful techniques include statistical method, such as Hidden Markov Models and probabilistic context-free grammars, and rule-based methods which employ some form of machine learning. One of the most successful machine learning method for IE is the Support Vector Machine (SVM), which is a general supervised machine learning algorithm and has achieved state-of-the-art performance on many classification tasks.

This paper will focus on information extraction for Public Health Surveillance. Injury surveillance systems based on traditional clinical derived data have the advantage of being a well established, highly reliable source of information for making an active surveillance (Rainey and Runyan 1992). However, there is a substantial lag time from the event to the reporting of such data and the sources are not routinely available to the public. Newspaper articles are cheap, available and have been reported to provide additional information about the events, especially when there is a death (Rainey and Runyan 1992; Ghaffar, Hyder et al. 2001). According to this alternative injury surveillance tool recognizing clinical words becomes critical and is important for information extraction (IE), in order to make organized and structured information available (Corney, Buxton et al. 2004; Zhou, Zhang et al. 2004).

Three different information extraction (IE) algorithms were implemented: a classical rule based system which requires a manual annotation of the rules of ; a rule based system which allows for the automatic building of rules; a machine learning method based on Support Vector Machine.

Results were compared by means of three performance measures, precision, recall and F-measure, with the aim of finding the best performed algorithm.

2. Materials and methods

2.1 Data Collection

A contract was established with a state-wide newspaper clipping service to obtain articles on both fatal and non-fatal children injuries due to suffocation. The clipping service provided newspaper articles describing injuries that occurred during 3 years since 2003 until September 2006. Articles were selected if the terms “soffocamento” and “bambini” (respectively suffocation and children) appeared in the text. Images of 388 typewritten text captured by scanner got from the newspaper clipping agency were turned into editable text using OCR tools. Finally, 44 newspaper articles were

selected and processed to extract information referred to age and gender of the injured child and to foreign body (FB) that caused the injury.

2.2 Information Extraction techniques

Two rule based systems and a support vector machine approach were used in the analysis to extract information from the newspaper articles. The first rule based system was a manual rules based approach built up by specifying a list of rules to be used for IE task, whereas in the second rule based system, rules were automatically detected through the definition of a dictionary that sufficiently covered domain information.

2.2.1 Rule based systems

A rule-based approach is a semi-automatic system which relies on a set of extraction rules provided by the user. Rule based Information Extraction was performed using GATE (General Architecture for Text Engineering) (Cunningham, Maynard et al. 2002) and VisualText (Text Analysis International Inc. 2001).

GATE is one of the most popular tools for Natural Language Processing. It includes a set of linguistic components shaped in an Information Extraction system called ANNIE (A Nearly-New Information Extraction system) which relies on the JAPE language.

All documents were pre-processed using ANNIE in order to obtain a number of linguistic features. The features included tokens, sentence splitting, and semantic classes from gazetteer lists. Firstly, ANNIE's tokenizer identified units of Natural Language such as words and character punctuation and treated each of the tokens as an annotation. Then a sentence splitter attempted to identify and annotate the beginning and the end boundaries of each sentence. New gazettes were added for children's gender, children's age and the type of foreign bodies that caused suffocation. A gazette is

a user-defined dictionary and consists of a list of single and multi-word encountered as entity. Finally, the Semantic tagger was run in order to produce annotated entities by a list of rules manually defined. Semantic tagger consists in a set of JAPE grammars, i.e. a set of rules, which acts on annotation assigned in earlier phases and produces outputs of annotated entities.

VisualText is based on NLP++ Programming language, which compiles to C++ and DLL files and it provided an automatic pattern approach. It worked on an alternative method to classical rule-based systems. Rules were determined automatically after the information to be extracted was supplied. Automatic acquisition of linguistic patterns partially performed this task. In order to make it possible, lexical resources were provided by the definition of a dictionary that sufficiently covered domain information.

2.2.2 Support Vector Machine system

By manually annotating a small number of documents with the information to be extracted, a reasonably accurate Information Extraction system can be induced from this labelled corpus and then applied to a large corpus of text to construct a database. In order to perform this task a Support Vector Machine (SVM) information extraction system was implemented using T-REX (Iria, Ireson et al. 2006).

Data were preprocessed using GATE which provided tokenization, Part-Of-Speech tagging and named-entity recognition text features. Moreover ANNIE's Semantic tagger was run in order to get linguistic features used in SVM input. Thus on a training set of 25 annotated documents an SVM-based classifier was performed and the learned model was tested on the remaining 19 articles. More specifically, the SVM implementation used in the experiment was SVM^{perf} (Joachims 2006). The default values of the parameters in SVM^{perf} were used.

2.3 Performance measures

Information Extraction systems were evaluated calculating Recall, Precision and the F-measure (Makhoul, Kubala et al. 1999). Only for the manual rule based system, three values were provided for each performance measure: a 'strict' performance values which considered the false positive answers as incorrect answers; a 'lenient' performance value which considered the false positive answers were considered as correct answers; an 'average' performance value which was the average between strict and lenient values.

3. Results

A set of 44 articles provided between 2003 and September 2006 were analyzed. An example of the information which had to be extracted was provided in Table 1, in particular a partial list of words used for detecting the foreign bodies was elaborated.

Judging by the performance measures, accuracy increased in the following order of techniques: support vector machine, automatic pattern rule base system and manual rule based system. The two rule based approaches performed essentially identically. Moreover the overall scores of the manual rule based system were 79.13% for Recall and 84.64% for Precision (see table 3). Further analysis of these results (Table 2), showed that performance measures are mainly influenced by the degree of correct extraction of gender and age information more than of the FB characteristics. On the contrary, while using the automatic rule based approach, we achieved regarding FB information a precision of 66.4% and a recall of 73.0%, resulting in a F-measure of 69.5%, that was better than the results regarding age information (details are showed in Table 4). Age information results with classical rule based system report an F-measure between 80.9% and 87.8% (see Table 2)

4. Conclusions

The growth and development of injury surveillance systems presents important challenges. With many sources that can provide data, there is a greater need for integration and methods for dealing with unconventional sources of data. In this sense, newspapers clippings provide information that is not always available from traditional public health datasets. For example, it makes possible to learn the circumstances surrounding the injury in which the child was involved such as the presence of other children or the parents, as reported in the journal and often not recorded in the hospital records.

In addition, data available from hospital discharge records and from mortality data in death certificates are coded using the International Classification of Diseases (ICD-9) indicating the nature of the injury, and, only when an injury is the result of an external cause, another code (E-code) is used in addition to the ICD code. This code defines both the manner of the injury and the mechanism of the event. However, coding of external causes has been considerably less complete for morbidity data, and this has limited the usefulness of sources such as hospital discharge records for injury surveillance and product safety assessment (Horan and Mallonee 2003). A great advantage of newspaper clippings is that, while available from such source, description of the event provided in articles could be useful to get the E-code. Indeed, as part of local sections, newspaper reports provide detailed information not just for fatal injuries.

For what concerns reliability of such data source, and in order to assess the usefulness of newspaper articles as a surveillance tool it should be linked to computerized state death and hospital records for determining what proportion of injuries were reported.

In this paper three IE techniques were analyzed: two rule based systems and a machine learning approach using SVM. Performance measures were higher using rule based systems than Support Vector Machine. As well as SVM has emerged as one of the leading trainable models for many classification task (Collier and Takeuchi 2004), the small number of articles related to choking injuries in children influenced the results. Automatic methods are less reliable but suitable for only

large volumes of articles (Ananiadou, Kell et al. 2006), rule based system perform well also working on a small text corpus (Marshall 2005) even if they are more time consuming.

Despite their intrinsic appeal, several limitations are still present in the widespread usage of newspaper clippings in surveillance systems: a major problem which arises in using newspaper clippings as data source is that many articles deal with the same injury. Furthermore on the same article more events could be reported. On the other ways, the process of reading, organizing and distributing media clips is usually time-consuming and expensive. Thus using non traditional source of data requires developing mathematical algorithms base on Inductive Text Analysis in order to get user's most relevant information and assemble a database for statistical analysis.

Nevertheless, the appeal in terms of timeliness and cost-efficacy of getting information from newspaper clipping is evident: research should be fostered in the direction of overcoming actual pitfalls in the methodology.

References

- Ananiadou, S., D. B. Kell, et al. (2006). "Text mining and its potential applications in systems biology." Trends Biotechnol.
- Baullinger, J., L. Quan, et al. (2001). "Use of Washington State newspapers for submersion injury surveillance." Inj Prev **7**(4): 339-42.
- Centers for Disease Control and Prevention (2001). Updated guidelines for evaluating public health surveillance systems: recommendations from the guidelines working group. MMWR Recomm Rep. **50**: 1-51.
- Collier, N. and K. Takeuchi (2004). "Comparison of character-level and part of speech features for name recognition in biomedical texts." J Biomed Inform **37**(6): 423-35.
- Corney, D. P., B. F. Buxton, et al. (2004). "BioRAT: extracting biological information from full-length papers." Bioinformatics **20**(17): 3206-13.
- Cunningham, H., D. Maynard, et al. (2002). GATE: a framework and graphical development environment for robust NLP tools and applications. 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02).
- Fine, P. R., C. S. Jones, et al. (1998). "Are newspapers a viable source for intentional injury surveillance data?" South Med J **91**(3): 234-42.
- Ghaffar, A., A. A. Hyder, et al. (2001). "Newspaper reports as a source for injury data in developing countries." Health Policy Plan **16**(3): 322-5.
- Guard, A. and S. S. Gallagher (2005). "Heat related deaths to young children in parked cars: an analysis of 171 fatalities in the United States, 1995-2002." Inj Prev **11**(1): 33-7.
- Horan, J. M. and S. Mallonee (2003). "Injury surveillance." Epidemiol Rev **25**: 24-42.

- Iria, J., N. Ireson, et al. (2006). An Experimental Study on Boundary Classification Algorithms for Information Extraction using SVM. Workshop on Adaptive Text Extraction and Mining 11th Conference of the European Chapter of the Association for Computational Linguistics.
- Joachims, T. (2006). Training Linear SVMs in Linear Time. , Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD).
- Makhoul, J., F. Kubala, et al. (1999). "Performance measures for information extraction." in Proceedings of DARPA Broadcast News Workshop, (Herndon, VA).
- Marshall, R. J. (2005). "Comparison of misclassification rates of search partition analysis and other classification methods." Stat Med **25**(22): 3787-3797.
- Rainey, D. Y. and C. W. Runyan (1992). "Newspapers: a source for injury surveillance?" Am J Public Health **82**(5): 745-6.
- Text Analysis International Inc. (2001). Integrated Development Environments for Natural Language Processing.
- Voight, B., G. Lapidus, et al. (1998). "Injury reporting in Connecticut newspapers." Inj Prev **4**(4): 292-4.
- Zhou, G., J. Zhang, et al. (2004). "Recognizing names in biomedical texts: a machine learning approach." Bioinformatics **20**(7): 1178-90.

Table 1 Partial list of words used for detecting the foreign body which caused the accident.

Foreign body type	
Italian word	English translation
pezzo di mandarino	mandarine orange
Pomodoro	tomato
Pallina	ball
pezzo di mela	piece of apple
Brioche	brioche
Palloncino	balloon
Moneta	coin
Batteria	battery
Carota	carrot
Mais	mais
Nocciolina	nut

Table 1. Performance measures (precision, recall and F-measure) calculated with regards to the type of information (gender, age and foreign body) for the manual rule based system only.

	precision	recall	F-measure	
FB	76.8%	69.5%	73.0%	average
	76.8%	69.5%	73.0%	lenient
	76.8%	69.5%	73.0%	strict
GENDER	89.9%	82.2%	85.9%	average
	89.9%	82.2%	85.9%	lenient
	89.9%	82.2%	85.9%	strict
AGE	74.4%	88.5%	80.9%	average
	77.3%	91.9%	84.0%	lenient
	71.6%	85.1%	77.8%	strict

Table 2: Precision, recall and F-measure performances calculated for each IE technique implemented (manual rule based system, pattern approach rule based system, Support Vector Machine).

		<i>performance measures</i>		
		<i>precision</i>	<i>recall</i>	<i>F-measure</i>
<i>IE system</i>	<i>Manual rule based</i>	84.64%	79.13%	81.79%
	<i>Pattern approach</i>	80.33%	80.33%	80.33%
	<i>Support Vector Machine</i>	33.33%	60.71%	43.04%

Table 3: Precision, recall and F-measure performances obtained according to the manual rule based system and the pattern approach to rule based system in detecting gender, age and foreign body information.

<i>information extracted</i>	<i>IE system</i>	<i>performance measures</i>		
		<i>precision</i>	<i>recall</i>	<i>F-measure</i>
<i>GENDER</i>	<i>Manual rule based</i>	89.94%	82.18%	85.89%
	<i>Pattern approach</i>	89.29%	99.43%	94.09%
<i>AGE</i>	<i>Manual rule based</i>	71.59%	85.14%	77.78%
	<i>Pattern approach</i>	79.07%	45.95%	58.12%
<i>FB</i>	<i>Manual rule based</i>	76.84%	69.52%	73.00%
	<i>Pattern approach</i>	66.39%	72.97%	69.53%

