Can ProMED-mail Bootstrap Blogs? Automatic Labeling of Victim-reporting Sentences*

Avaré Stewart and Kerstin Denecke

L3S Research Center
Appelstr. 9A, 30169 Hannover, Germany

Abstract. Due to the proliferation of social media data and user-generated content available, monitoring trends or using this data in other scenarios becomes more interesting. Our research focuses on the extraction of information on health events from user generated content with the objective to support Epidemic Intelligence. Specifically, we describe and evaluate a method for identifying sentences relevant for event extraction. Labeled data is unavailable for this task and manual annotation is expensive. Therefore, in order to reduce the number of labeled examples, we apply a bootstrapping algorithm for this task. In more detail, we will study the suitability of a classifier trained on one text type (e-mails) for the classification of texts of another text type (blogs).

1 Introduction

The spread of infectious diseases and - due to this - the increased public concern - raises the necessity to have health surveillance systems on hand for detecting disease outbreaks as early as possible. All the activities related to early identification of potential health hazards, their verification, assessment and investigation with the objective to recommend public health control measures are summarized by the term *Epidemic Intelligence* [1]. A *public health event* is an event that creates a need for action of public health officials, for instance an outbreak of an infectious disease or one case of a very seldom infectious disease. A public health event can be described by *event information* providing information on *who* was infected by *what*, *where* and *when*, i.e., information on a victim, a location, a time or a disease.

Besides the traditional surveillance systems that monitor indicators such as death rates, drug prescriptions, occurrence of viruses etc. event-based systems have been developed. They extract and analyse outbreak-related information from text in electronic sources such as e-mail, official reports, news wires and present the results to the user. Social media data or user-generated content (e.g., Weblogs, Twitter messages) remained so far unconsidered for Epidemic Intelligence. In our research, we will focus on this text type.

The problem of detecting health events can be decomposed into mainly three sub-problems:

* This work has been done within the M-Eco project, partly funded by the European Commission under 247829.

1. Annotation: Identifying sentences containing information on an event.
2. Event Extraction: Identifying relevant facts to describe the event.
3. Event Aggregation: Aggregating information on the same event that has been reported in different sources.

In this paper, the focus is on identifying pieces of text relevant for public health event detection (Annotation problem). State of the art approaches for detecting public health events either rely upon a huge number of extraction patterns or a large set of labeled data in order to detect the relevant information-bearing sentences. However, when extracting such information specifically from social media, additional challenges are faced which make these approaches inadequate[2] including use of specific language, and different styles of writing. Large amounts of social media data are available. The data is noisy to a large extent; and it is often opinionated and can contain irrelevant information.

We address these challenges using semi-supervised learning in the context of event extraction. In particular, the sub-problem of identifying sentences relevant to the detection of health events on infectious diseases is considered. We will study the suitability of a classifier trained on one text type for the classification of texts of another. In more detail, patterns acquired from a manually curated source are transferred to bootstrap the classifier for medical blogs. To reduce manual work for labeling training data, we come up with an approach of automatically labeling a training set that bases upon research results known from the field of text summarization. The automatically determined training set is then used to learn a classifier for sentence classification.

The contributions of this work are (1) presentation of a health event detection framework, (2) introduction of a cross-corpus bootstrapping framework to identify relevant sentences, and (3) study related corpora for the task of bootstrapping.

2 Related Work

Our final goal is to extract public health events for outbreak detection. Existing systems for this task rely upon the enumeration of possible types of victim reporting patterns (e.g., MediSys [3], HealthMap [4], BioCaster [5]). MediSys for example uses manually specified keywords in different languages to identify news articles reporting on health events. In these systems, little or no attention is devoted to using blogs and other forms of user generated content as source of information. Other systems rely upon the linguistic, interpretive, and analytical expertise of analysts to filter and extract information about health threats (e.g., GPHIN [6]). Given the dynamic nature of social media in general and of weblogs in particular, a pattern-based system is not realizable since the number of patterns needed may be numerous. Further, the large amount of user generated content available requires application of automatic methods. In this paper, an approach based on semi-supervised learning for identifying relevant sentences for event extraction is introduced.

Bootstrapping approaches are such semi-supervised learning approaches and can be grouped into self-training and co-training approaches. Semi-supervised self-training has been applied in the field of sentiment analysis (e.g., subjectivity analysis [7]). Didaci and Role [8] compare several semi-supervised learning methods for multiple classifier systems. Chen and Ji [9] present a framework where bootstrapping is used for event extraction in a cross-lingual setting. An event extraction system in one language is bootstrapped by exploring new evidences from a system in another language. In this work, we apply an existing bootstrapping algorithm to the task of sentence classification. To the best of our knowledge, this particular task of classifying sentences of user generated content for health event detection has not been considered before. Further, bootstrapping has not been used in this context before. We will study the quality of such approach and report the results.

Methods which make use of unannotated text and intra-document information have emerged as important approaches for information gathering. The systems typically rely upon the redundancy present on the web, and assume that facts with multiple mentions are more reliable [3]. Zhen and Li [10] consider the problem of cross-domain text classification in the news domain. They propose a support vector based semi-supervised algorithm to solve this problem. Yangarber et al. [11] use cross-document analysis to support building a consistent and robust fact base about epidemics or the outbreak of disease. In our work, the cross-corpora analysis is applied to detect victim-reporting sentences. Cross-classification is applied to train a classifier on a data set of an auxiliary domain to classify data of the target domain. Besides reporting quality results of the approach, we will study the conditions when it is possible to use such a learner.

3 Approach

The objective of our approach is to identify information bearing sentences in social media, when faced with the problem of large amounts of unlabeled data. In particular, we consider a sentence relevant when it contains information on disease outbreaks (see section 4 for more details).

Given the characteristics of social media data, a supervised classification approach seems to be better suited than pattern-based approaches that rely upon extensive manual work. Our objective is to reduce the manual work for labeling examples as much as possible. For this reason, we make use of data of an auxiliary domain to determine labeled training examples. In more detail, for an auxiliary domain, we build a classifier in a bootstrap process. This classifier is then applied to label sentences of the target domain. To avoid manual labeling, we introduce an approach to automatic labeling examples of the auxiliary domain. The single processing steps are described in more detail in the next sections.

3.1 Automatic Labeling

For learning the classifier, a related and complementary, auxiliary data source is used. This related source typically uses a terse and more compact style of prose. It

also exhibits structural properties which allow us to apply a weak form of labeling - i.e., automatically label selected sentences as positive or negative examples with respect to disease reporting based on their position in the document. This is in contrast to the sentences in the target domain, for which we have a less obvious structural pattern from which we can weakly label sentences.

In more detail, the idea is to use sentences at the beginning of a document as positive examples. This idea has already been proven successful in the field of text summarization where sentences at the beginning of a document are used to produce a document summary.

Let $D = d_1, d_2, \ldots, d_j$ be the set of documents in an auxiliary corpus, where each document, $d_i \in D$, consists of a set of one or more sentences. Further, let $T = t_1, t_2, ..., t_m$ be a set of feature types used to represent the sentences of D. We kept the approach general with respect to the set of features to be used. Types of features can include bag-of-words, bag-of-concepts, part of speech, or a typed dependency structure.

Given a set of documents D and a surrogate representation for a given type, T_t applied to the sentences in D, a corpus can be modeled as a sentence database, $S_t = s_{t11}, s_{t12}, ..., s_{tjk}$, where s_{tjk} represents the k^{th} sentence for the j^{th} document using feature type, t. Further, we label the top-N sentences in the database automatically as positive cases and the bottom-N as negative examples, for a threshold value of N.

At this stage, we also want to introduce the concept of *sublanguages*. Given a sentence database of type t, S_t, for an auxiliary domain. Then, we can define the auxiliary corpus to be a sublanguage for the target corpus if a self-trained classifier built from the auxiliary domain performs well on the unlabeled examples in the target domain with some threshold tolerance. In our experiments, we will study whether the dataset from the auxiliary domain is a sublanguage for the target corpus.

3.2 Cross-Corpora Bootstrapping

Using the previously described approach to automatically labeling the sentences of the auxiliary domain, we determine a set of labeled examples to be used to train a classification model using bootstrapping. We are considering these examples produced by the automatic labelling approach as *weakly labeled*, i.e., there is some confidence why they have been selected as positive or negative examples, but it is unclear to what extent this labeling is confident. To reduce bias produced by this incertitude, we produce an improved classifier through a bootstrap process. The bootstrap process is depicted in Figure 2. The algorithm is described in more detail in the following (see Figure 1).

3.3 Classifying Sentences

The previous step produces a classifier trained on material of the auxiliary domain. We have chosen Support Vector Machines as classification algorithm. Finally, this classifier is applied to the sentences of the target domain and labels

Input:
W: a set of weakly labeled examples from an auxiliary domain
L: set of labeled examples from the auxiliary domain
C_i: a set of classifiers

Initialization:
Initialize the set L by selecting the weakly labeled candidates from W. Create a pool W_{prime} of examples by choosing p examples from W

Loop until a condition is satisfied (e.g., W = $\emptyset$, or threshold number of iterations reached)

- Train each classifier C_i on L, and label the examples in W_{prime}
- For each classifier C_i, select the most confidently labeled examples (e.g., the confidence score is above an absolute or relative threshold for the current run) and add them to L
- Refill W_{prime} with examples from W, and keep the size of W_{prime} as constant p

Fig. 1. Bootstapping Algorithm

the sentences of the target domain as positive or negative, or victim-related and not victim-related, respectively.

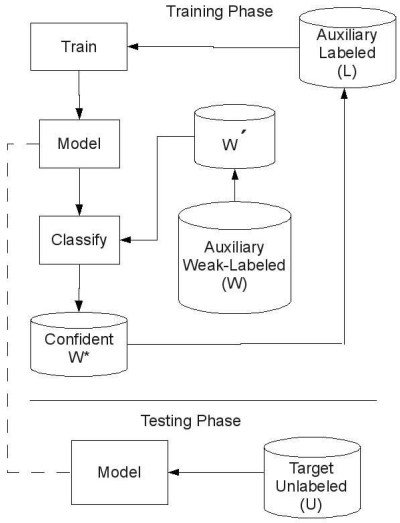


Fig. 2. Weakly Labeled Bootstrap: Overview

4 Experiments

In this preliminary work the experimental goals are twofold. First, we are interested in knowing how good is a classifier based on weak labeling at identifying sentences bearing information on public health events in blog posts. Second, we are interested in characterizing the conditions under which the auxiliary source can be considered a sublanguage for the noisier corpus. We characterize noise based on the length of the sentences or the type of entities appearing in them. We

conduct experiments from two perspectives of the bootstrapping process, namely training and testing phases. Experimental setting and results are described in the next section. At the end, we will discuss the approach.

4.1 Experimental Setting

In our experiments, we use data collected from ProMED-mail [12] as the auxiliary domain. ProMED-Mail (referred to as Promed) is a global electronic reporting system, which lists outbreak reports of emerging infectious diseases. It constitutes a terse source of information about epidemic events. Similarly, the World Health Organization also reports disease outbreak news on their webpage (`http://www.who.int/csr/don/en/index.html`). This data is considered as yet another moderated data source, referred to as WHO.

The data of the target domain is provided by the AvianFluDiary (`http://afludiary.blogspot.com/`). All data was collected directly from the websites. Summary statistics for the data are shown in Table 1.

Source	Years	No. of Documents	No. of Sentences
AvianFluDiary	2006-2009	4249	100890
ProMed-Mail	2002-2009	13369	22170
WHO	1996-2009	1531	16213

Table 1. Data Collection for Experiments

The goal of our experiments is to identify information bearing sentences in AvianFluDiary blogs. As can be seen in Table 1, even for a single blog, spanning less than half the number of years for each moderated source, the number of blog sentences is still over three times greater. We therefore seek to evaluate, the effectiveness of a weakly labeled classifier at detecting relevant disease reporting sentences in such a voluminous and more verbose data source.

In order to do so, we train a classifier on data material of our auxiliary domain (Promed) based on the structural SVM implementation of SVM-TK v1.2 [13]. The features used in the classifier are the tree structure of the parts of speech for each sentence. To create these features, the text was normalized to remove extraneous symbols. Sentence splitting and parse trees were created by the Stanford Parser. Named entities were recognized by applying OpenCalais (`http://www.opencalais.com/`).

As initial training documents for training the classifier, we created a weakly labeled set of examples for the auxilary domain using the automatic labeling approach introduced in section 3.1. In more detail, the top-1 sentences in the sentence database were labeled as positive cases and the bottom-1 as negative examples for our classification task. Observing that not all positive and negative sentences were of equal quality, we prefiltered all top-1 sentences based on

the sentence length. We used default values where the minimum and maximum sentence lengths were set to 20 and 200, respectively.

To test the classifier, a total of 5,029 (729 positive and 4,599 negative) AvianFluDiary sentences were hand labeled. The sentences were labeled with respect to the task of identifying *victim reporting* sentences. The definition we used for victim reporting is based on the MedISys Disease Incidents template[1]. The template reports *disease, time, location, status, cases, description* and *url* to a natural language text.

Therefore, we label a sentence as victim reporting sentence if it contains: D = {disease} in union with I = {time, location, status, cases}. *Status* reports the condition of a victim (e.g., *hospitalized, dead*), and *cases* refers to the number and type of victims (e.g., *bird, child*, etc). Further, we labeled all sentences needed to report a single event as a positive case with the value 1; all other sentences are labeled with 0.

4.2 Experimental Results

The objective of the experiments were two-fold: Determining the quality of the introduced classification approach (Part I), and characterizing conditions for sublanguages (Part II).

Part I: Bootstrapping with Automatically Labeled Data For the first part of the experiments we were interested in determining the quality of a classifier at identifying information bearing sentences when it is trained on weak labeled training material of an auxiliary domain. We tested the classifier on the AvianFlu-Diary data. The bootstrap learner takes into account different scenarios of the bootstrap process.

- **Scenario 1:** Default settings based on the values used by the authors for a similar task (auxiliary domain: Promed),
- **Scenario 2:** Applied on bottom-1 sentences only that are additionally filtered (auxiliary domain: Promed),
- **Scenario 3:** Scenario 1 with WHO as auxiliary data,
- **Scenario 4:** Sentences are filtered based on presence of named entities (auxiliary domain: Promed)

In Scenario 1, the pool size was set to 15, and 50 sentences per iteration was used. A stopping condition was reached when 2,000 items in the weakly labels set were labeled (model size). A classified sentence was selected as confident if its confidence value exceeded 70% percent of the maximum confidence value relative to the given iteration. The initial pool size of 50 positive and 50 negatives sentences was used.

In Scenario 2, similar parameters are used, except that the model size is reduced to 1700. In Scenario 3, WHO data is used as an auxiliary source of data,

[1] `http://medusa.jrc.it/medisys/helsinkiedition/all/home.html`

to see if applying another weakly labeled sentence database yield to different results. Finally, in Scenario 4, we filter the sentences based on the presence of named entities which contain both a medical condition and location. For the different scenarios, the precision, recall and accuracy values of the learner on the AvianFlu-Diary data is examined (see Table 2).

Scenario	Precision	Recall	Accuracy
1	.77	.45	.57
2	.71	.66	.69
3	.75	.22	.34
4	.80	.40	.53

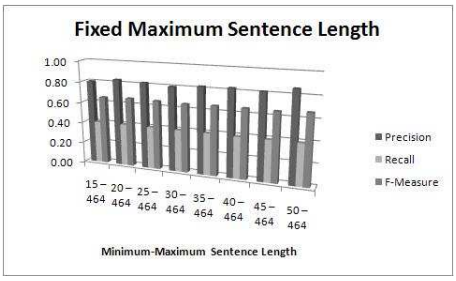
Table 2. Bootstrap Results per Scenario

It can be seen that for the different scenarios differing accuracy values are achieved. The best accuracy of .69 is determined for scenario 2, while the worst accuracy of .34 is achieved for scenario 3. Precision values lie between .71 and .8 for all four scenarios. The recall is significantly lower with values between .22 and .66. We discuss these results in Section 4.3.

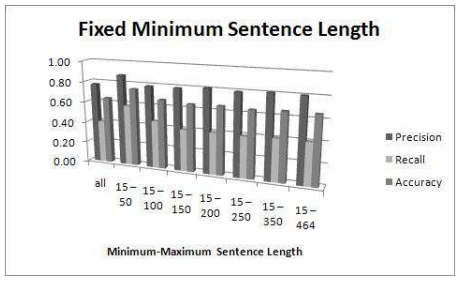
Part II: Analysing Sublanguage Conditions In the second part of experiments, we are interested in characterizing the conditions under which the language of an auxiliary data source can be considered a sublanguage for the noisier corpora. We account for noise by filtering the length of the sentences in the target data and filter based on the type of entities appearing in the sentences. The training sentence lengths were filtered with a minimum sentence length of 30 and a maximum sentence length of 200. The results are shown in Figures 3. When varying the minimum sentence length, the best precision of more than 80% is achieved for sentences with a minimum length of 50. For shorter sentences, the precision drops below 80%. Further, when varying the maximum length of sentences, the best results are achieved for a maximum sentence length of fifty words. In the next section, these results will be interpreted and discussed.

4.3 Discussion

In part I of the experiments (see section 4.2), an important observation is that given a threshold value of .7 for precision, ProMed-Mail can in fact be used in an automatic way to build a bootstrap classifier for blogs. This implies that by using the top-1 sentences as positive cases we are capable of identifying relevant health related sentences in blog postings. Given the fact to no human effort was incurred for labeling a training set, a fairly good classifier can be built based on weak labeling for identifying sentences bearing information on public health events.



(a) Fixed Maximum Sentence Length



(b) Fixed Minimum Sentence Length

Fig. 3. Part II: Analysis of Sublanguage Conditions

We also notice that except for Scenario 2, in which additional filtering was applied to the bottom-1 sentences, the recall tends to be quite low. This would suggest that although top-1 positive cases perform well, the bottom-1 negative examples used in training are not representative enough to distinguish the negative examples present in the blogs. In light of this, we propose that further experiments are needed to better characterize the conditions under which the auxiliary source can be considered a sublanguage for the noisier corpora for identifying negative cases. In part II of the experiments 4.2, we seek notice that although for most of the different sentence lengths, the same results in precision and recall are achieved, in a single range for both the fixed upper and fixed lower, we achieve a noticeable peak precision above 80%. This suggests that such an approach is sensitive to the length of sentences in the test data.

In summary, the introduced approach for sentence classification has been proven to provide acceptable results. In contrast to existing approaches, the main benefits of the method presented here are: (a) Avoidance of manual labeling of training material, and (b) Reduction of bias produced by automatic labeling through bootstrapping for learning the classifier. The presented approach is new for several reasons: Bootstrapping has not yet been applied for learning a classifier in a cross-corpora setting for labeling sentences. The problem of victim-reporting sentences classification was only considered using pattern-based approaches by now. We introduce a weakly-supervised approach to address this problem. Further, the automatic labeling of examples and its combination with bootstrapping to reduce incertitude has not been reported and analysed before.

5 Conclusion

In this work, an approach is described to identify disease- and victim-reporting sentences from blogs to support epidemic intelligence. Challenges given by the characteristics of blogs (mainly noise and data abundance) are addressed using automatic labeling of data collected from moderated sources to learn a classifier for identifying relevant sentences present in blogs. A bootstrap process is applied

to learn a classifier and filter more noisy and irrelevant sentences. The results show that the approach taken here is quite effective at sentence level filtering in blogs. Without manual effort, we are able to achieve a precision as high as .80 and a recall .66. In the future, we will perform robust experiments, particularly using more blogs. We also intend to experiment with increasing values for top-N and compare this to bootstrapping on the blog set alone and using a hybrid approach.

References

1. C. Paquet, D. Coulombier, R.K., Ciotti, M.: Epidemic intelligence: A new framework for strenghtening disease surveillance in europe. Euro Surveill. **11(12)** (2006)
2. Moens, M.F.: Information extraction from blogs. In Jansen, B.J., Spink, A., Taksa, I., eds.: Handbook of Research on Web Log Analysis, IGI Global (2009) 469–487
3. Yangarber, R.: Verification of facts across document boundaries. In Proceedings International Workshop on Intelligent Information Access (2006)
4. Freifeld, C.F., Mandl, K.D., Reis, B.Y., Brownstein, J.S.: Healthmap: Global infectious disease monitoring through automated classification and visualization of internet media reports. In Proceedings International Workshop on Intelligent Information Access (2006)
5. Collier, N., et al.: Biocaster: detecting public health rumors with a web-based text mining system. Bioinformatics, Oxford University Press (2008)
6. Mykhalovskiy, E., Weir, L.: The global public health intelligence network and early warning outbreak detection: a canadian contribution to global public health. Can J Public Health **97(1)** (2006) 42–44
7. Wang, B., Spencer, B., Ling, C., Zhang, H.: Semi-supervised self-training for sentence subjectivity classification. Canadian AI 2008, LNAI 5032 (2008) 344–55
8. Didaci, L., Roli, F.: Using co-training and self-training in semi-supervised multiple classifier systems. In: D.-Y. Yeung et al. (Eds.): SSPR&SPR 2006, LNCS 4109 (2006) 522–530
9. Chen, Z., Ji, H.: Can one language bootstrap the other: a case study on event extraction. In: SemiSupLearn '09: Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing, Morristown, NJ, USA, Association for Computational Linguistics (2009) 66–74
10. Zhen, Y., Li, C.: Cross-domain knowledge transfer using semi-supervised classification. In: AI '08: Proceedings of the 21st Australian Joint Conference on Artificial Intelligence, Berlin, Heidelberg, Springer-Verlag (2008) 362–371
11. Yangarber, R., et al.: Combining information about epidemic threats from multiple sources. In: RANLP-2007, Borovets, Bulgaria (2007)
12. Madoff, L.C.: Promed-mail: An early warning system for emerging disease. Clinical Infectious Diseases **2**(39) (July 2004) 227–232
13. Moschitti, A.: A study on convolution kernels for shallow semantic parsing. In: ACL '04, Morristown, NJ, USA, Association for Computational Linguistics (2004) 335