Reducing Noise in Labels and Features for a Real World Dataset: Application of NLP Corpus Annotation Methods

Rebecca J. Passonneau*, Cynthia Rudin†, Axinia Radeva‡, and Zhi An Liu§

Columbia University, New York, NY 10027, USA
*becky@cs.columbia.edu, (†cr2363|§zl2153)@columbia.edu,
‡axinia@hotmail.com

Abstract. This paper illustrates how a combination of information extraction, machine learning, and NLP corpus annotation practice was applied to a problem of ranking vulnerability of structures (service boxes, manholes) in the Manhattan electrical grid. By adapting NLP corpus annotation methods to the task of knowledge transfer from domain experts, we compensated for the lack of operational definitions of components of the model, such as *serious event*. The machine learning depended on the ticket classes, but it was not the end goal. Rather, our rule-based document classification determines both the labels of examples and their feature representations. Changes in our classification of events led to improvements in our model, as reflected in the AUC scores for the full ranked list of over 51K structures. The improvements for the very top of the ranked list, which is of most importance for prioritizing work on the electrical grid, affected one in every four or five structures.

1 Introduction

This paper illustrates how a combination of information extraction, machine learning, and NLP corpus annotation techniques was applied to a problem of ranking vulnerability of structures (manholes and service boxes) in the Manhattan electrical grid. Institutions of all sorts collect and archive large amounts of data. The value of the data to the institution depends in part on whether methods can be developed to make use of it. Information extraction, defined as the task of organizing and normalizing data taken from unstructured text in order to populate tables in structured databases, has obvious relevance, as does machine learning. Automated techniques for corpus annotation clearly support the task of information extraction. What is less obvious, and potentially of greater impact, is that the practices developed in the NLP community for manual annotation and classification of documents provide an effective means to arrive at clearer problem definitions. By adapting NLP corpus annotation methods to the task of knowledge transfer from domain experts, we compensated for the lack of operational definitions of components of the model, such as *serious event*.

During the 1970s, the Consolidated Edison Company instituted a program of Emergency Control System (ECS) tickets to document calls from customers about potential problems in the electrical grid. Beginning in 1986, after Hurrican Gloria, the ECS program expanded and became more fully utilized. As the

archive grew, Con Ed hypothesized that information in the ECS tickets could be used to rank the vulnerability of manholes and other structures in the network to serious events. We worked with Con Ed to refine and test this hypothesis.

A discussion of related work is in the next section, followed by a section presenting background on the scope of the problem, and an example of an ECS ticket for a moderately serious event. The next two sections describe the information extraction (section 4) and machine learning (section 5). In section 6, we present the human annotation task, and the impact on our labeling and feature representation of examples. Section 7 presents the results of a comparison of alternative labelings: a baseline, our initial rule-based labeling, and the rule-based labeling that we now use, derived from the results of the human annotation task. In the final section we discuss the implications and conclusions.

2 Related Work

Machine learning has been used to predict failures in the primary (high voltage) network for Consolidated Edison [1], but not using data extracted from free text fields. Con Ed trouble tickets [2], maintenance logs for complex machinery [3], naval equipment reports [4], aeronautic safety reports [5] and similar sets of documents have been handled using the methods of natural language processing, knowledge modeling, and machine learning for a wide range of goals. Relatively early work [4] showed it was possible to handle fragmentary text using full syntactic and semantic parsing, but involved a much smaller dataset than current work on ticket databases. Devaney and Ram [3] deal with 10,000 logs, all of which pertain to the same machines. They combine unsupervised text clustering with a domain representation modeled in OWL/RDF to classify tickets, then develop a Case-Based Reasoning approach to predict failures. Liddy and her colleagues [2] developed an application for the same type of trouble ticket data we address, but their goal was to assign ECS tickets with a miscellaneous categorization to a more specific ticket type.

Oza [5] is the only work we have seen that deals with a similarly large dataset, and where disagreements among human experts made it difficult to define document classes. They look at two aeronautics report databases that have a combined size of 800,000 reports. Their reports, unlike ECS tickets, generally have a single author, and consist of a readable, discursive narrative. Their end goal is to arrive at a comprehensive, topic-based document classification, whereas our classification task is to scale the severity of events, and we ignore ticket content not relevant to this task. They rely on an existing thesaurus (PLADS) to merge distinct forms of a single term, such as acronyms, abbreviations and phrases, making their documents amenable to a bag-of-words (BOW) document representation, and they use two learning techniques, Support Vector Machines and Non-negative Matrix Factorization. Our early attempts to use BOW features foundered due to the high noise content. In ongoing work, we have been looking at decision trees for document classification, and clustering methods to generate string normalization rules.

3 Background and Example

We have 1,036,732 ECS tickets for 1996 through 2006, from four New York City boroughs. They fall into hundreds of distinct *trouble types*, a ticket category Con Ed assigns to each ECS ticket. The first line of the ticket in Figure 1 explicitly names the category for this ticket, which is SMH (smoking manhole); MHF is for a manhole fire, LV is for a low voltage problem, and so on. For Manhattan, we investigated twenty-two trouble types, comprising 61,730 tickets for the ten-year period. A ticket represents a report about an event or problem that the caller judged to involve the electrical grid. These data are noisy: there are far more tickets than there are distinct *events*, and many more distinct events than those we use for labeling and feature representation. Furthermore, not all tickets mention a specific structure (manhole or service box), and they are not intended to contain a complete description of every event.

1 01/21/*YR* 18:45 FDNY-190 REPORTS A SMH *STREET_1* & *STREET_2*
2 01/21/*YR* 19:35 *PERSON* REPORTS THE **TROUBLE HOLE** IS **SB-00001**
3 N/W/C *STREET_1* & *STREET_2*.....FOUND ON ....SMOKING LIGHTY
4 01/21/*YR* 21:55 *PERSON* REPORTS IN **SB-00001** HE FOUND 1 LEG
5 ON THE 5 WIRE NORTH BURNING IN THE STRUCTURE.....CUT/CLEAR
6 ED & RETIED SAME .....COMPLETE.....SS
7 ELIN REPT ADDED FOR INCIDENT:SMH 01/21/*YR* 22:02BY *PERSON_ID*
8 REPORTED BY: FIRE DEPT
9 STRUC MSPLATE TYPE NUMBER COND COVTYP COVFOUND DISTANCE

Fig. 1. Sample ECS ticket (anonymized): serious smoker

To use the ECS ticket data, we addressed four tasks: identifying structures mentioned in the ECS tickets, pruning the tickets to a set of unique events relevant for our modeling task, labeling the events as serious or not, and deriving ECS-based and other features to represent structures. The final three tasks depended heavily on the outcome of the human annotation task.

One of the questions we faced at the outset was what time frame we could make predictions about. In separate work [6], we report exploratory data analysis indicating that, given the data made available to us, we could make predictions based on longer term hotspots. The Manhattan machine learning model described there, and presented to Con Ed, ranks structures for a given one-year period, based on data from prior years. The ranking criterion is the likelihood that a structure will experience a serious event within the current year.

Counts of structures and events give a sense of the scope of our task. There are 51,219 structures in Manhattan. In the 61K tickets we investigate, we extract mentions of 27,235 structures (see section 4). Depending on the definition of *serious event*, we estimate that there are on the order of seventy-five to five hundred serious events per year in Manhattan. Defining more precisely what counts as a "serious" event is the focus of this work. The trouble type of a ticket is a good but not perfect indicator of seriousness for three trouble types

(MHX, MHF, MHO), and a moderate indicator for a fourth (SMH). In our most recent rule-based classification of events into serious and non-serious, which we know overgenerates, there are 5,115 serious events in Manhattan for the ten year period. If a ticket is one of the most serious trouble types–MHF (fires), MHX (explosions) and MHO (open manholes, or possible explosions)–chances are we classify it as serious (1,481 out of 1,506 tickets, or 98.3%). The additional 3,634 serious events include 3,397 SMHs (smoking manholes)–a somewhat less serious trouble type that we classify as serious about 75% of the time, ACBs (AC burnounts, N=162)–which we classify as serious 2.9% of the time, and a mixed set of thirteen trouble types (N=75). Here we use these six categories of trouble types: MHX, MHF, MHO, SMH, ACB, all others.

ECS tickets can have multiple entries from different individuals, some of which is free text, some of which is automatically entered. In Figure 1, a slightly modified version of a relatively readable ticket that we currently classify as serious, we show only the free text lines we are concerned with here. ECS tickets exhibit the fragmentary language, lack of punctuation, acronyms and special symbols characteristic of trouble tickets [4]. They have a high rate of misspellings (see "lighty" in line 12, a typical misspelling) and line breaks within words (see "clear {newline} ed" in lines 6-7). Evidence to classify this event as serious is the mention of a smoking manhole in line 1 and the degree of smoke in line 3. Evidence that in another context could point to a non-serious event is that the smoking manhole is the report of someone other than an engineer (line 1), and that the cover is on ("found on," line 3). The structure is named as the "trouble hole" (line 2), a domain expression meaning the event in question occurred in this structure. Tickets rarely identify the trouble hole explicitly, and often mention multiple structures.

4 Information Extraction from ECS Tickets

The ECS tickets in our subset range in length from 1 to 550 lines, with a similarly wide range of information content. The information we extract falls into two categories. One category corresponds to domain-specific named entities, where a text string names an object in the domain. The entities we extract consist of cables and structures. Structures have a type (service box versus manhole) and a numeric identifier; see the service box (SB 00001) in Figure 1. The combination of type, number and location provides a triple that is used in retrieving the unique identifier for the structure from Con Ed's asset table. Our structure extraction achieves about 90% accuracy against a known subset [6]. We do not use existing named entity recognizers because the types of entities we identify are domain-specific, and because the text is too divergent from standard orthography (e.g., a high degree of misspellings, acronyms, non-standard symbols, and words with linebreaks within them).

Given a structure involved in an event, the two most critical questions for our task are, was work performed on the structure, and was the structure involved

in a serious event. How we addressed the second question is the subject of this paper, and is described in section 6.

5 A Ranking Approach to the Learning Task

The goal of our collaborative effort with Con Ed is to produce a ranking of structures according to vulnerability to serious manhole events. We formulated the task as a *supervised bipartite ranking* problem. Within this framework, machine learning algorithms are used to provide a real-valued score for each structure. It requires a set of *examples* with *labels*. The goal is to construct a model that ranks the positively-labeled examples above the negatively labeled examples, and this model should generalize to other (non-labeled) examples chosen from the same probability distribution.

In this domain, a structure changes over time as events occur, insulation breaks down, cables are repaired or replaced. Thus our examples consist of structures paired with a given time frame, which in our case is a year. A structure gets a positive label if it was the trouble hole of a serious event during the relevant year (Y_i), and a negative label otherwise.

A structure is represented by a set of features that characterize the structure before the year Y_i from which the model is built. Figure 2 lists the features we use for the learning models in this paper. The cable feature (F5) is one of several alternate features we have used to capture density of cables in a structure.

	ECS-based features
F1	number of times structure is a trouble hole between the start of 1996 and Y_i
F2	number of ECS tickets mentioning the structure between the start of 1996 and Y_i
F3	number of times structure is a trouble hole in the 3 year period before Y_i
F4	number of ECS tickets mentioning the structure in the 3 year period before Y_i
	Other feature: cable-based
F5	number of neutral mains cables in the structure

Fig. 2. Five features (F1-F5) used to represent structures

Much of our effort has been devoted to developing an accurate, streamlined, interpretable and intuitive model, as described in [6]. Four of five features pertain to the events a structure has been involved in. These ECS-based features play a key role for the top of the ranked list, but provide little information regarding mid- or low-ranked structures; conversely, the cable feature (F5) has little impact at the top of the ranking, but plays a large role for mid-ranked structures. A central issue for our work is that the labels and ECS-based features both depend on our ability to identify relevant events, and to classify them as serious or not.

We construct a training set (Y_i=2005) and a test set (Y_i=2006). We determine whether the model is statistically predictive by constructing the model from the training set and using it to predict the labels in the test set. Consider the structure SB-00001 mentioned in the ticket in Figure 1, which was classified

as a serious ticket. If the ticket date is in 2006, then this ticket pertains to the structure's label in the test data. If the ticket date is in 2005, the ticket pertains instead to the label of SB-00001 in the training data, and to its feature representation in the test data. In the latter case, the structure is the trouble hole of the event, so the values of F1-F4 would all be incremented.

The machine learning algorithm in essence maximizes a proxy for the AUC (area under the Receiver Operator Characteristic curve), or a weighted version of the AUC, which can be viewed as a measure of ranking quality. Machine learning algorithms for this task include RankBoost [7] the P-Norm Push [8] and IR-Push (used in [6]) and SVM-perf [9], used in this work.

6 Human Annotation Task

There are three decisions we make for each ECS ticket that enable us to label and represent examples. First, we must decide if the ticket documents a distinct event. Multiple customers might call to report the same event, generating a new ticket for each call, but with the repair work on one ticket that the other tickets will cross-reference. We refer to the latter category as *referred* tickets, and filter them from the dataset. Second, we must decide if the documented event is relevant to the status of the secondary electrical grid (as opposed to the higher voltage primary grid). Third, for each relevant event, we must decide whether it is serious. The annotation task pertains to decisions two and three.

The three domain experts we consulted with had far more than adequate expertise to interpret tickets for us. One of them, who is now a manager, was the programmer and representative to code maps and input ECS databases for 1985-1991. We gradually realized that due to the variation in tickets, the complexity of the domain, and time limitations, we would never acquire sufficient domain expertise through interviews. An even greater obstacle was a lack of operational definitions of relevant, serious and non-serious events. Approximately eighteen months into the project, we are still learning constraints on relevant events, such as the fact that larger buildings occasionally have 265 volt service, but with no connection to the secondary grid (120 volt).

6.1 Assembling and Assessing the Annotations

Human annotations of natural language data are collected for a wide range of purposes, although the most common one today is probably to assemble training and testing data for supervised machine learning tasks. The goal is often to use machine learning to replicate directly a human classification task. Our goal is to use the experts' annotations of trouble tickets to develop a manually derived rule-based classifier, and subsequently to use the features we derive from the trouble tickets to support our distinct machine learning task, namely ranking the vulnerability of structures within a given time frame. The main obstacle we faced in also developing a machine learning approach for classifying the tickets is that we lacked the domain knowledge to create training data ourselves, and

lacked access to experts' time. As we describe below, we were able to show significant gains in our machine learning task by extrapolating rules to classify tickets from a small set of 171 expert-annotated tickets.

One hundred and seventy one tickets were selected for the annotation task in a fashion that biased the set in two ways, but that otherwise aimed for an unbiased selection. First, we created a bias towards a somewhat higher proportion of serious events than in a random subset. The motivation was that after filtering out referred tickets, the proportion of the most serious trouble types (MHX, MHF, MHO) is relatively low (4% for the ten year period versus 14% if SMH is included; 1.6% or 15% for 2005, 2.6% or 12% for 2006), possibly too low to provide a sufficiently general set of examples, given the relatively small dataset that the experts agreed to label. Second, we created a bias towards a higher proportion of tickets in locations that experienced a series of events within a few months. Here the motivation was that we had observed that a structure was more likely to experience a serious event if it had already had a history of serious or non-serious events (see [6]). The way we created this bias was that we first generated ticket histories consisting of a ticket, along with all tickets for locations within a sixty meter radius of the original ticket that occurred in the preceding two month period, and the first ticket for the same location that occurred in the following month, if there was one. We eliminated tickets whose histories contained no serious trouble types. Then we made a random selection of one hundred histories. The final set of histories contained one hundred seventy-one tickets of assorted trouble types, with approximately 20% being serious.

Two experts agreed to do the annotation task. To test whether each expert had self-agreement, four of the tickets were repeated at random positions within a scrambled ordering of the set of tickets. Each was given the tickets in a different order. The two experts worked independently. They had no access to the trouble type, and were asked not to consult any other data sources. The annotation guidelines consisted of one page of instructions asking the experts to classify each ticket into exactly one category: serious, not serious, or not relevant.

Interannotator agreement (IA) coefficients measure agreement above chance, using the formula below, where $p(A_O)$ is the proportion of observed agreement, and $p(A_E)$ of expected agreement.

$$\frac{p(A_O) - p(A_E)}{1 - p(A_E)} \tag{1}$$

Agreement coefficients differ in how to estimate the probability that annotators will agree [10]. The NLP community often relies on Cohen's κ [11] or Krippendorff's α [12]. In practice, their values are often quite close. How to interpret IA values is the subject of much debate (see review in [10]). For Landis and Koch [13] in the medical arena, values between 0.40 and 0.60 are considered moderate; Krippendorff [12] recommends that for Content Analysis, values above 0.67 support *cautious conclusions* and lower values are suspect.

Trouble type is not a good means of classifying the seriousness of events: IA between a baseline classification based on trouble type and the expert consensus

(for the tickets where there is consensus) is poor (κ=0.25; α=0.20). Table 1 gives a breakdown of the 171 tickets. It shows that experts identified 29 serious tickets compared with a baseline of 36, 82 precursor tickets compared with a baseline of 127, 21 irrelevant tickets versus 8, and disagreed on 39 (22.8%). In thirteen of the disagreements, one expert classified the ticket as serious; we asked them to resolve the disagreements, yielding four more serious tickets.

	Serious		Non-serious		Not relevant		
	Baseline	Expert	Baseline	Expert	Baseline	Expert	Disagree
ACB	0	3	21	16	0	0	2
MHO	2	2	0	0	0	1	0
MHF	5	3	1	0	0	1	0
MHX	2	2	0	0	0	0	0
SMH	27	15	0	7	0	3	2
Other	0	4	106	58	8	17	35
Totals	36	29	127	82	8	21	39

Table 1. Human classification of 171 events

IA is measured on the 171 unique tickets, using the experts' first response to the one ticket where they disagreed with themselves. Here, κ is 0.4863 and α is 0.4865, or about halfway between 0 (chance distribution) and 1 (perfect agreement). For the four tickets that were randomly duplicated, experts agreed with themselves three of four times.

The moderate IA values between the experts indicates they agree well above chance, yet there is also a fair degree of subjectivity; data from different annotators might lead to different results for some cases. We concluded from the poor performance of the baseline classification of events against the yardstick of expert judgements that we needed an alternative classification of ECS events, despite the relatively modest IA between the experts. We had already begun developing a rule-based approach to sorting tickets into the three categories, based on the domain knowledge that we were slowly acquiring. To show the evolution of our event classification, we use three alternative classification methods: a baseline based on trouble type and a single length constraint (at least three lines per ticket), our pre-annotation rule-based method (Rules1), and the rule-based method we developed based on the annotation task (Rules2).

6.2 Reducing Noise in Labels and Features

We aimed to improve the precision of event classification overall, and to improve the recall of serious events. For tickets both experts agreed on, we used regularities observed within the three classes of tickets to hypothesize constraints to add to our classification rules. We applied the constraints to our database of tickets, manually evaluated random samples, then reiterated until the results of random sampling appeared to meet our two goals. The changes to our rules affected all three classes: non-relevant events, serious events, and precursor events.

Table 2 shows changes in the distribution of serious versus non-serious tickets by trouble type across the three classification methods. The most dramatic changes are the reduction in the number of relevant tickets, which reduces the size of both the non-serious and serious classes by 26.5%, and the big shift in the classification of SMHs from 100% serious down to 75.5% serious. The number of relevant events decreased from the baseline of 38,911 to 28,987 in the first rule based approach (Rules1), to 28,587 in the final rule-based approach (Rules2). Tickets were dropped on the basis of several criteria, including the voltage value (see above), and changes to length constraints on several categories of tickets.

	Non-Serious			Serious		
Type	Baseline	Rules1	Rules2	Baseline	Rules1	Rules2
MHX	0	2	0	179	177	160
MHF	0	79	21	1011	931	829
MHO	0	38	4	595	549	492
SMH	0	584	1105	4906	4284	3397
ACB	6171	4594	5364	192	582	162
Other	25776	14967	16978	81	2210	75
Totals	31947	20264	23472	6964	8733	5115

Table 2. Changes in the distribution of serious versus non-serious tickets

The number of precursor events decreased from 31,947 in the baseline to 20,264 in Rules1, then increased to 23,472 in Rules2. The percentage of all relevant events that are precursors decreased to 69.91% in Rules1 from 82.10% in the baseline, and increased back to the same percentage of 82.11% in Rules2.

The reduction in the number of SMH tickets classified as serious involved manual identification of a dictionary of phrases indicative of serious events, such as "manhole explosion," which is unambiguous but rare, or "mh smoking heavy" which has many variants (including "smoking lightly"), and other phrases with the words "fire" and "blown." For each *dictionary entry*, we identified variant patterns, and relevant contexts, including contexts with negation.

7 Results

To highlight the comparison of the three methods, we distinguish between structures that have or have not been mentioned in any ECS tickets prior to Y_i. In the testing data (2006), there are 21,471 structures that have been mentioned in ECS tickets from prior years back through 1996, versus 29,748 that have not been mentioned. Using the baseline classification of events, the proportion of positively labeled instances in the mentioned structures is $\frac{232}{21,471}$, and in the not-mentioned structures it is $\frac{197}{29,748}$. This gives a ratio of 1.6317, which is well above random labeling (a ratio of 1). Using Rules1, the proportion of positively labeled structures that have been mentioned to positively labeled structures that have

not been, is 1.6788. For Rules2, the proportion of positively labeled structures in the mentioned set is 1.9602, or almost double the baseline.

	Train	Test
Baseline	67.63	65.01
Rules1	66.80	63.72
Rules2	68.29	67.55

Table 3. AUC scores for the three event classification methods

Table 3 shows the AUC values for the three methods. Rules2 exhibits the least difference between training and test, and has the highest values. Due to many factors, for instance the high skew in the data, and the fact that information derived from ECS tickets primarily affects the top of the ranked list of structures, the changes to the event classification that reduce the noise in labels and features do not lead to a dramatic difference in AUC values. The impact of the improvements shows up in a more qualitative analysis of the ranked lists.

To illustrate the change in quality of the rankings, we present details on two structures that shifted rank, and which exemplify the changes in the ranked list. We will refer to the two structures as D (for demotion) and P (for promotion). The largest shifts in position were demotions, and D illustrates why. Using the baseline classification, structure D was mentioned in eleven precursor tickets, and was the trouble hole five times. In contrast, using our Rules2 event classification, D had only eight precursor tickets, and was the trouble hole twice. It changed in rank from 759/38911 (at 1.95% from the top) in the baseline, to 3823/28997 (13.18%) in Rules1, to 3105/28587 (10.86%) in Rules2. The structure's label does not change, but the ECS features shift in the direction of lower vulnerability.

In contrast, the promotion of structure P does not reflect changes in the feature representation of P; rather, it is a side-effect of the demotions that occurred around it. In the baseline feature representation, P was a trouble hole seven times and mentioned in twelve tickets, versus five times a trouble hole out of nine mentions in Rules1, and back to seven times a trouble hole in Rules2, out of eleven mentions. It changed from rank 69/38911 (0.18%) in the baseline to 205/28997 (0.71%) in Rules1 to 45/28587 in Rules2 (0.16%).

The demotion of structures that are not as serious as they might appear, and the correlated promotion of structures that are genuinely serious, is exactly the type of change we hoped for, and that makes it easier for Con Ed to fold our results into the way they prioritize structure repairs. Small improvements in AUC scores are not relevant to Con Ed unless they can also see changes in the ranked list, such as a higher proportion of obviously problematic structures.

Tables 4-5 highlight differences in the ranked lists. Table 4 shows the Jaccard distance [14] between the top N structures of the Rules2 ranked list versus the baseline's, where we look at various values of N. For two sets, Jaccard is the ratio of the size of set intersection to the size of the set union, thus is closer to 0 when the sets have fewer members in common and is 1 when the two sets are identical. From the AUC values in Table 3, we know that Rules2 is more predictive than

N	5	10	15	20	25	100	300	500	1000	2500	5000	10000
Jaccard	0.25	0.33	0.50	0.60	0.56	0.72	0.74	0.75	0.81	0.81	0.86	0.89

Table 4. Jaccard distance of Rules 2 from the baseline on the test data for the top N of the ranked lists

Rules1 or the baseline, but the Jaccard values tell us more specifically that the highest ranked structures in the Rules2 ranking are most different up to the top N structures, with only one a quarter to a third of the structures in common, and that even at N=500, every fourth structure in Rules2 is not included in the baseline ranking (or every fifth structure for N=2500). For ConEd, the top of the list has a different status; all structures are inspected on a five year cycle, but they would like to identify a subset that needs attention first.

N	Baseline	Rules1	Rules2
5	36.40	35.80	48.30
10	46.40	42.00	48.30
100	46.49	46.07	46.42
500	41.60	41.58	41.34
5000	35.08	35.61	35.11
50000	25.34	25.34	25.34

Table 5. Average vulnerability on the test data for the top N of the ranked lists

As an investigative tool, we had earlier constructed a vulnerability score (a non-statistical measure) for structures, in close consultation with domain experts; the features it uses do not help the learning, but do help explain the ranking results. It ranges from 0 to 100, with scores above 65 being rare. Table 5 shows that for the top N structures in the ranked lists from the learning models, the average vulnerability per structure decreases from about 50 for the top 100 to about 25 for the full list. Rules 2 gives the highest average vulnerability for the top 10 structures, and about the same as the baseline for the rest of the list. Thus Rules2 pushes more highly vulnerable structures to the very top of the list.

8 Discussion and Conclusion

Our modeling task depended heavily on information found in an extremely noisy and disparate set of documents. Our document classification determined both the labels of examples and their feature representations, an interdependence not often found in learning tasks addressed within NLP. Machine learning has become a very powerful tool, but successful learning requires a knowledge transfer from the human to the machine. In NLP, this often takes the form of human annotations on documents, which serve either as the relevant labels for a learning task, or the features. Since both labels and features depended on the document classification, our model is doubly sensitive to it. The experts' annotations

helped us propose manually derived classification rules: our human intelligence became the mechanism of knowledge transfer from the experts to the learning. As a result, the model is more predictive, as described in section 7. The AUC, which we are using as a general measure of ranking quality, does not reflect the dramatic improvement at the top of the ranked list. One out of four or five structures in the top 500-2500 of our ranking was not present in the corresponding part of the baseline ranking. Also, a non-predictive but interpretable scoring of vulnerability shows that the top 5 to 10 structures in the Rules2 ranking have a higher average vulnerability score. Thus the new ranking is qualitatively better than the AUC scores indicate on their own.

References

1. Gross, P., Boulanger, A., Arias, M., Waltz, D.L., Long, P.M., Lawson, C., Anderson, R., Koenig, M., Mastrocinque, M., Fairechio, W., Johnson, J.A., Lee, S., Doherty, F., Kressner, A.: Predicting electricity distribution feeder failures using machine learning susceptibility analysis. In: The 18th Conference on Innovative Applications of Artificial Intelligence IAAI-06, Boston, Massachusetts (2006)
2. Liddy, E.D., Symonenko, S., Rowe, S.: Sublanguage analysis applied to trouble tickets. In: Proceedings of the Florida Artificial Intelligence Research Society Conference. (2006) 752–757
3. Devaney, M., Ram, A.: Preventing failures by mining maintenance logs with case-based reasoning. In: Proceedings of the 59th Meeting of the Society for Machinery Failure Prevention Technology (MFPT-59). (2005)
4. Hirschman, L., Palmer, M., Dowding, J., Dahl, D., Linebarger, M., Passonneau, R., Land, F., Ball, C., Weir, C.: The PUNDIT natural-language processing system. In: Proceedings of the Annual AI Systems in Government Conference. (1989) 234–243
5. Oza, N., Castle, J.P., Stutz, J.: Classification of aeronautics system health and safety documents. IEEE Transactions on Systems, Man and Cybernetics, Part C (Accepted for publication)
6. Rudin, C., Passonneau, R.J., Radeva, A., Dutta, H., Ierome, S., Isaac, D.: Predicting vulnerability to serious manhole events in manhattan: A preliminary machine learning approach Submitted for publication.
7. Freund, Y., Iyer, R., Schapire, R.E., Singer, Y.: An efficient boosting algorithm for combining preferences. Journal of Machine Learning Research **4** (2003) 933–969
8. Rudin, C.: The P-Norm Push: A simple convex ranking algorithm that concentrates at the top of the list. Accepted, Journal of Machine Learning Research (2008)
9. Joachims, T.: A support vector method for multivariate performance measures. In: Proceedings of the Internat'l Conf. on Machine Learning (ICML). (2005) 377–384
10. Artstein, R., Poesio, M.: Inter-coder agreement for computational linguistics. Computational Linguistics (To appear)
11. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement **20** (1960) 37–46
12. Krippendorff, K.: Content analysis: An introduction to its methodology. Sage Publications, Beverly Hills, CA (1980)
13. Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33(1)** (1977) 159–174
14. Jaccard, P.: Nouvelles recherches sur la distribution florale. Bulletin de la Societe Vaudoise des Sciences Naturelles **44** (1908) 223–270