Semi-Supervised Cause Identification from Aviation Safety Reports

Isaac Persing and Vincent Ng
Human Language Technology Research Institute
University of Texas at Dallas

Project Mission

Develop a set of data mining and text analysis methods for systematic, multidimensional analysis of the ASRS database

Aviation Safety Reporting System

- established in 1967
- voluntarily submitted reports about aviation safety incidents written by flight crews, attendants, controllers, …

Cause Identification

- determines **why** the incident described in a report occurred
- Not a sentence or phrase extraction task
- A **text categorization** task
  - Experts at NASA have identified 14 causes (or *shaping factors*) that could explain why an incident occurred
  - **Goal**: given an incident report, determine which of a set of 14 shapers contributed to the occurrence of the incident

Shaping Factors (Posse et al., 2005)

- **Proficiency**
  - general deficit in capabilities
    - inexperience, lack of training, not qualified, …
- **Attitude**
  - unprofessional attitude by a controller or flight crew member
    - complacency, in a hurry to get home, …
- **Physical Factors**
  - pilot ailment that could impair flying
    - being tired, drugged, ill, dizzy, …

Shaping Factors (Cont')

- **Physical Environment**
  - physical conditions that could impair flying
    - snow, hurricane, …
- **Communication Environment**
  - interferences with communications in the cockpit
    - noise, auditory interference, radio frequency congestion, …

Shaping Factors (Cont')

- **Resource Deficiency**
  - absence, insufficient number, or poor quality of a resource
    - overworked or unavailable controller, insufficient or out-of-date chart, malfunctioning or missing equipment
- **Unexpected**
  - something sudden and surprising that is not expected
- **Other**
  - anything else that could be a shaper
    - shift change, passenger discomfort, disorientation, …
- **Familiarity**, **Pressure**, **Preoccupation**, **Taskload**, **Duty Cycle, Illusion**

Cause Identification is Challenging

- **No publicly available labeled data**
  - NASA researchers hand-annotated 20 reports only
- **Multi-label categorization**
  - an incident may be caused by more than one factor
  - categories not mutually exclusive
- **Skewed class distributions**
  - some shapers occur a lot more frequently than the others
  - some shapers cover a broad range of issues
  - 10 of the 14 shapers are minority classes
- **May require a deeper understanding of the text than topic-based classification**

Goal

- Improve cause identification
  - via a **bootstrapping** algorithm that augments a training set
    - learning from **labeled** data and **unlabeled** data
  - focus on improving **minority** class prediction

Plan for the Talk

- Dataset
  - Preprocessing, human annotation
- Two baseline cause identification methods
- Our bootstrapping algorithm
- Evaluation

Plan for the Talk

- Dataset
  - Preprocessing, human annotation
- Two baseline cause identification methods
- Our bootstrapping algorithm
- Evaluation

Dataset

- ~140K aviation safety reports in the ASRS database
- Each report is a free narrative, describing
  - why the incident happened
  - what happened
  - where the incident happened
  - how the reporter felt about the incident
  - the reporter's opinions of other people involved in the incident

Lots of information irrelevant to cause identification

Data Preprocessing

- Reports are informally written
  - domain-specific abbreviations and acronyms
  - poor grammar
  - capitalization information removed

- Example sentence

  “HAD BEEN CLRED FOR APCH BY ZOA AND HAD BEEN HANDED OFF TO SANTA ROSA TWR.”

Data Preprocessing

- Reports are informally written
  - domain-specific abbreviations and acronyms
  - poor grammar
  - no capitalization information

- Example sentence

  "HAD BEEN CLRED FOR APCH BY ZOA AND HAD BEEN HANDED OFF TO SANTA ROSA TWR."

  - Grammatically incorrect
  - Many abbreviations and acronyms

Three Preprocessing Steps

1. Dictionary-based acronym and abbreviation expansion
   - list taken from NASA's website
2. Heuristic-based case restoration
3. Stemming

Human Annotation

- Randomly picked 1,333 preprocessed reports
- Two graduate students annotated them with shapers
  - based solely on the definition of the shapers
  - Kappa value: 0.45
    - task is difficult
    - definition is vague
- Same two annotators re-examined each report for which there was a disagreement and reach an agreement

Dataset Statistics

Resource Deficiency	30.0
Physical Environment	16.0
Proficiency	14.4
Other	13.3
Preoccupation	6.7
Communication Environment	5.5
Familiarity	3.2
Attitude	2.4
Physical Factors	2.2
Taskload	1.9
Pressure	1.8
Duty Cycle	1.8
Unexpected	0.6
Illusion	0.1

Dataset Statistics

Resource Deficiency	30.0
Physical Environment	16.0
Proficiency	14.4
Other	13.3
Preoccupation	6.7
Communication Environment	5.5
Familiarity	3.2
Attitude	2.4
Physical Factors	2.2
Taskload	1.9
Pressure	1.8
Duty Cycle	1.8
Unexpected	0.6
Illusion	0.1

Dataset Statistics

Resource Deficiency	30.0
Physical Environment	16.0
Proficiency	14.4
Other	13.3
Preoccupation	6.7
Communication Environment	5.5
Familiarity	3.2
Attitude	2.4
Physical Factors	2.2
Taskload	1.9
Pressure	1.8
Duty Cycle	1.8
Unexpected	0.6
Illusion	0.1

Minority Shapers

Dataset Statistics (Con't)

- Percentage of reports with *n* labels

1	53.6
2	33.2
3	10.3
4	2.7
5	0.2
6	0.1

Plan for the Talk

- Dataset
  - Preprocessing, human annotation
- Two baseline cause identification methods
- Our bootstrapping algorithm
- Evaluation

Baseline Approaches

- Hypothesis
  - our bootstrapping algorithm for augmenting the labeled data can improve performance of the cause identification task
- Baselines
  - learn from labeled data only
  - recast problem as a set of 14 binary classification tasks
    - train one classifier for predicting whether a report has a particular shaper or not

Learning the Binary Classification Tasks

- Goal: train a classifier c_i for identifying shaper factor s_i
- Training data creation ("one versus all" method)
  - create one training instance from each training document
  - label the instance as
    - positive if document has s_i as one of its labels
    - negative otherwise
- Features
  - Top 50 unigrams selected according to information gain
- Learning algorithm
  - LIBSVM

Baseline 1

- All learning parameters are set to their default values
- 5-fold cross validation
- Results in terms of overall recall, precision, F1

$$\text{Recall} = \frac{\sum_i \text{no. of reports correctly labeled as positive for shaper i}}{\sum_i \text{no. of positive reports w.r.t. shaper i in gold standard}}$$

$$\text{Precision} = \frac{\sum_i \text{no. of reports correctly labeled as positive for shaper i}}{\sum_i \text{no. of reports labeled as positive by classifier i}}$$

Baseline 2

- Similar to Baseline 1, except that we tune the classification threshold (CT) to optimize F-measure
- Motivation
  - a classifier trained by LIBSVM by default employs a CT of 0.5
    - instance is classified as positive if and only if CT at least 0.5
  - may not be optimal threshold, especially for minority classes
- 5-fold CV:1 fold for tuning, 3 folds for training, 1 fold for testing
- 14 CTs are jointly tuned to optimize F-measure
  - computationally expensive
  - employ a local search algorithm that alters one parameter at a time and holds the remaining parameters fixed

Plan for the Talk

- Dataset
  - Preprocessing, human annotation
- Two baseline cause identification methods
- Our bootstrapping algorithm
- Evaluation

Our Bootstrapping Algorithm

- Goal
  - Improve the baseline classifiers by training them on training data augmented using the bootstrapping algorithm

Idea

- Given a training set created for shaper *s*, iteratively
  - identify words that are strong indicators of the positive or negative examples of shaper *s*
  - automatically label unlabeled documents that contain a sufficient number of such indicators

Mutually bootstrap the feature set and the labeled data

Algorithm for augmenting training data for shaper *s*

- Input arguments
  - L^+: set of positively labeled training examples of shaper *s*
  - L^-: set of negatively labeled training examples of shaper *s*
  - *U*: set of unlabeled documents
  - *k*: number of bootstrapping iterations
- Variables
  - W^+: words that are strong indicators of positive examples
  - W^-: words that are strong indicators of negative examples

Repeat for *k* iterations
  if $|L^+| > |L^-|$
    Expand L^- and W^-
  else
    Expand L^+ and W^+

Expand the smaller of L^+ and L^-

Algorithm for augmenting training data for shaper *s*

- Input arguments
  - L^+: set of positively labeled training examples of shaper *s*
  - L^-: set of negatively labeled training examples of shaper *s*
  - *U*: set of unlabeled documents
  - *k*: number of bootstrapping iterations
- Variables
  - W^+: words that are strong indicators of positive examples
  - W^-: words that are strong indicators of negative examples

Repeat for *k* iterations
  if $|L^+| > |L^-|$
    Expand L^- and W^-
  else
    Expand L^+ and W^+

Expand the smaller of L^+ and L^-

Expanding L^+ and W^+

1. Find the four words in the labeled data ($L^+ \cup L^-$) that are the strongest indicators of the positive examples according to the log likelihood ratio

$$LL(w) = \frac{\text{(number of reports in } L^+ \text{ containing } w)}{\text{(number of reports in } L^- \text{ containing } w) + 1}$$

2. Expand W^+ with these four words

3. Label all documents in U containing at least 3 words in W^+ as positive and add them to L^+

Expanding L^+ and W^+

1. Find the four words in the labeled data ($L^+ \cup L^-$) that are the strongest indicators of the positive examples according to the log likelihood ratio

$$LL(w) = \frac{\text{(number of reports in } L^+ \text{ containing } w)}{\text{(number of reports in } L^- \text{ containing } w) + 1}$$

?

2. Want to prevent the algorithm from selecting words that appear frequently in L^+ and not at all in L^-

3. Label all documents in U containing at least 3 words in W^+ as positive and add them to L^+

Expanding L^+ and W^+

1. Find the four words in the labeled data ($L^+ \cup L^-$) that are the strongest indicators of the positive examples according to the log likelihood ratio

   Want to ensure with a reasonable level of confidence that the newly added documents should be labeled as positive

2. Expand W^+ with these four words

   ?

3. Label all documents in U containing at least 3 words in W^+ as positive and add them to L^+

Expanding L^+ and W^+

?

1. Find the four words in the labeled data ($L^+ \cup L^-$) that are the strongest indicators of the positive examples according to the log likelihood ratio

Want to prevent the algorithm from selecting words that are too specific to one subcategory of a shaping factor
(e.g., for Physical Environment, after choosing "snow", "plow" will more likely to be chosen than "hot")

2.

3. Label all documents in U containing at least 3 words in W^+ as positive and add them to L^+

Number of Bootstrapping Iterations

- Between 0 and 5
  - decided against running for more than 5 iterations, as the quality of bootstrapped data deteriorates rapidly
- to be tuned on held-out data

Plan for the Talk

- Dataset
  - Preprocessing, human annotation
- Two baseline cause identification methods
- Our bootstrapping algorithm
- Evaluation

Evaluation

- Goal
  - evaluate the effectiveness of bootstrapping (with a focus on minority class prediction)
    - determine whether the baseline classifiers can be improved when trained on the augmented training data
- 5-fold cross validation
  - results are micro-averaged over the five folds

Baselines

- train 14 SVM classifiers, one for predicting each shaper
- Baseline 1
  - uses default values for all learning parameters
  - 4 folds for classifier training, 1 fold for testing
- Baseline 2
  - tunable parameters are the 14 CTs from the 14 classifiers
  - allowable values for each CT are 0.1, 0.2, …, 1.0
  - jointly tuned to optimize F-measure on held-out data
  - 3 folds for classifier training, 1 fold for tuning, 1 fold for testing

Results (Baseline 1)

	All 14 Classes			Minority Classes		
	R	P	F	R	P	F
Baseline 1 ($B_{0.5}$)	34.4	67.0	**45.4**	23.9	68.3	**35.4**

Results (Baseline 1)

	All 14 Classes			Minority Classes		
	R	P	F	R	P	F
Baseline 1 ($B_{0.5}$)	34.4	67.0	**45.4**	23.9	68.3	**35.4**

Results (Baseline 1)

	All 14 Classes			Minority Classes		
	R	P	F	R	P	F
Baseline 1 ($B_{0.5}$)	34.4	67.0	**45.4**	23.9	68.3	**35.4**

Results (Baseline 1)

	All 14 Classes			Minority Classes		
	R	P	F	R	P	F
Baseline 1 ($B_{0.5}$)	34.4	67.0	**45.4**	23.9	68.3	**35.4**

Results (Baseline 1)

	All 14 Classes			Minority Classes		
	R	P	F	R	P	F
Baseline 1 ($B_{0.5}$)	34.4	67.0	**45.4**	23.9	68.3	**35.4**

Results (Baseline 1)

	All 14 Classes			Minority Classes		
	R	P	F	R	P	F
Baseline 1 ($B_{0.5}$)	34.4	67.0	**45.4**	23.9	68.3	**35.4**

- F-measure using all 14 shapers are higher than using 10 shapers
  - due to improvements in recall
  - small number of positive instances for minority classes, yielding a bias towards classifying an instance as negative

Results (Baseline 2)

	All 14 Classes			Minority Classes		
	R	P	F	R	P	F
Baseline 1 ($B_{0.5}$)	34.4	67.0	**45.4**	23.9	68.3	**35.4**
Baseline 2 (B_{CT})	59.2	47.4	**52.7**	34.3	47.8	**39.9**

- In comparison to the Baseline 1
  - F-measure rises by 7.4% (14 shapers) and 4.5% (10 shapers)
- Employing the right CT is important

Bootstrapping Experiments

- Train the baselines, $\mathbf{B}_{0.5}$ and $\mathbf{B}_{CT}$, on the **expanded** training data to produce two systems, $\mathbf{E}_{0.5}$ and $\mathbf{E}_{CT}$, respectively

- For both systems, *k* (number of iterations) is a tunable parameter with allowable values ranging from 0 to 5
  - $\mathbf{E}_{0.5}$: *k* is the only parameter to be tuned
    - the 14 values of *k* tuned jointly using a local search algorithm
  - $\mathbf{E}_{CT}$: both *k* and *CT* need to be tuned
    - use local search
    - in each search step, adjust both *k* and *CT* for exactly one of the 14 classifiers to optimize overall F-score

Results (Bootstrapping Experiments)

	All 14 Classes			Minority Classes		
	R	P	F	R	P	F
Baseline 1 ($B_{0.5}$)	34.4	67.0	**45.4**	23.9	68.3	**35.4**
Baseline 2 (B_{CT})	59.2	47.4	**52.7**	34.3	47.8	**39.9**
$E_{0.5}$	40.4	60.9	**48.6**	35.3	53.2	**42.4**

Results (Bootstrapping Experiments)

	All 14 Classes			Minority Classes		
	R	P	F	R	P	F
Baseline 1 ($B_{0.5}$)	34.4	67.0	**45.4**	23.9	68.3	**35.4**
Baseline 2 (B_{CT})	59.2	47.4	**52.7**	34.3	47.8	**39.9**
$E_{0.5}$	40.4	60.9	**48.6**	35.3	53.2	**42.4**

- In comparison to Baseline 1 ($B_{0.5}$)
  - F-measure rises by 3.2% (14 shapers) and 7.0% (10 shapers)
  - due to a large gain in recall and a smaller drop in precision
    - recall can be improved with a larger training set
    - precision can be hampered when learning from noisy data
- Learning from augmented training set is useful, especially for minority classes

Results (Bootstrapping Experiments)

	All 14 Classes			Minority Classes		
	R	P	F	R	P	F
Baseline 1 ($B_{0.5}$)	34.4	67.0	**45.4**	23.9	68.3	**35.4**
Baseline 2 (B_{CT})	59.2	47.4	**52.7**	34.3	47.8	**39.9**
$E_{0.5}$	40.4	60.9	**48.6**	35.3	53.2	**42.4**
E_{CT}	54.9	50.5	**52.6**	39.4	49.1	**43.7**

Results (Bootstrapping Experiments)

	All 14 Classes			Minority Classes		
	R	P	F	R	P	F
Baseline 1 ($B_{0.5}$)	34.4	67.0	**45.4**	23.9	68.3	**35.4**
Baseline 2 (B_{CT})	59.2	47.4	**52.7**	34.3	47.8	**39.9**
$E_{0.5}$	40.4	60.9	**48.6**	35.3	53.2	**42.4**
E_{CT}	54.9	50.5	**52.6**	39.4	49.1	**43.7**

- In comparison to Baseline 2 (B_{CT})
  - F-measure drops by 0.1% for 14 shapers but rises by 3.8% for 10 shapers
- When CT is tunable, bootstrapping helps minority classes but hurts the remaining classes

Results (Bootstrapping Experiments)

	All 14 Classes			Minority Classes		
	R	P	F	R	P	F
Baseline 1 ($B_{0.5}$)	34.4	67.0	**45.4**	23.9	68.3	**35.4**
Baseline 2 (B_{CT})	59.2	47.4	**52.7**	34.3	47.8	**39.9**
$E_{0.5}$	40.4	60.9	**48.6**	35.3	53.2	**42.4**
E_{CT}	54.9	50.5	**52.6**	39.4	49.1	**43.7**

- For the 4 non-minority classes, slight drop in F-measure
  - due to a large drop in recall and a smaller gain in precision
  - automatically labeled data either provides little new knowledge or are too noisy to be useful
  - decent classifiers can be trained using the original labeled data

Results (Bootstrapping Experiments)

	All 14 Classes			Minority Classes		
	R	P	F	R	P	F
Baseline 1 ($B_{0.5}$)	34.4	67.0	**45.4**	23.9	68.3	**35.4**
Baseline 2 (B_{CT})	59.2	47.4	**52.7**	34.3	47.8	**39.9**
$E_{0.5}$	40.4	60.9	**48.6**	35.3	53.2	**42.4**
E_{CT}	54.9	50.5	**52.6**	39.4	49.1	**43.7**

- For the 10 minority classes, gain in F-measure
  - due to a simultaneous gain in recall and precision
  - bootstrapped documents have provided useful knowledge, particularly in the form of positive examples, for the classifiers
  - classifiers trained on the original training data were not good, as the number of positive examples is typically too small

Summary

- Introduced a new problem: cause identification
- Hand-annotated 1,333 reports with shaping factors; see http://www.hlt.utdallas.edu/~persingq/asrsDataset.html
- Presented a bootstrapping algorithm for improving minority classes in the presence of a small training set