Subjectivity Classification and Analysis of the ASRS Corpus

Jason Switzer, Latifur Khan, etc
Department of Computer Science, University of Texas at Dallas
jswitzer@gmail.com, lkhan@utdallas.edu, etc

Abstract

Semantic analysis of corpora containing heavy usage of jargon words and phrases introduces problems not commonly addressed by Natural Language Processing methods. Modern semantic analysis relies on data from unedited websites or other expertly written sources, which lack similar usage of jargon words and phrases. This paper presents a system of semi-supervised lexicon learning algorithms that collate several manually labeled and clustered data sources, such as thesauri. In addition, this paper demonstrates an improvement in performance of these subjectivity classifiers by applying a boosting method. This paper presents a method of automatic Aviation Safety Reporting System (ASRS) shaping factor classification based on the most relevant words from a subjectivity lexicon.

Keywords: Classification, Sentiment, Text mining, Annotation

1. Introduction

An impetus behind causal analysis of the Aviation Safety Reporting System, or similar corpora, involves identifying the underlying cause of the report in spite of the existence of jargon. An automatic approach to classifying an individual report (or sentence) is the most desirable method of causal analysis.

There has been an abundance of recent research focused in sentiment analysis, which is an area of focus in the field of Natural Language Processing (NLP). Sentiment analysis, or subjectivity classification, involves classifying a word or phrase as positive or negative in most contexts. Subjectivity classification has applications to problems such as market analysis, public opinion, and customer feedback (Mohammad, Dunne, and Dorr 2009).

The Aviation Safety Reporting System corpus, ASRS, contains voluntarily submitted aviation incident reports. NASA builds the ASRS corpus from incidents reported by pilots, air traffic controllers, dispatchers, flight attendants, and mechanics at the rate of approximately 5,000 reports per month (Reynard et al. 1986). Given the source of the reports, the terminology frequently used in each report often contains jargon words or phrases specific to the aviation system. The high occurrence of jargon presents issues not present with other corpora, such as the New York Times Annotated Corpus (Sandhaus 2008), a heavily edited and highly regarded corpus compiled from the New York Times newspaper.

The reports in the ASRS corpus typically refer to deficiencies in the aviation system. As such, there should be a number of words and phrases used in negative context. The subjective words in the reports either are direct causes or surround the root cause of an incident. We present a technique of automatic causal analysis based on the set of nearby subjective phrases.

We demonstrate a system of automatic classification of the ASRS **shaping factors**, which indicate the cause of an incident. We will build a subjectivity lexicon by combining inferred knowledge from several English reference sources. Lastly, we will demonstrate that while the subjectivity lexicon is simple in approach, it is capable of automatically identifying phrases indicative of the shaping factors. This paper will focus on building a subjectivity lexicon, which we will use to build an entropy classifier.

We will begin by discussing recent related work. Then, we will discuss the ASRS corpus and its significance. Afterwards, we will detail the subjectivity classification system and classifiers. Finally, we will discuss the results and their meaning.

2. Related Work

The primary basis of this research is a paper by Mohammad et al. to develop an automatic classification scheme based on the Macquarie Thesaurus (Mohammad, Dunne, and Dorr 2009). They developed a set of affix seed patterns to mark antonymous word pairs and trained a subjectivity lexicon based on those 11 patterns. They also proposed a means to expand the subjectivity lexicon by using thesaurus entries for discovery and classification. While the performance reported in the paper is impressive, this paper will demonstrate even further improvements, both in the size of the trained lexicon and the accuracy.

(Esuli and Sebastiani 2006) developed a method to assign semantic orientation labels to WordNet synsets, known as SentiWordNet. Their technique augments

WordNet by training a random walking semi-supervised classifier. While this approach is similar to certain aspects presented here, it has deficiencies that this paper aims to rectify. Their random walk phase uses only the gloss information for a word, whereas WordNet 3.0 also provides synonyms that indicate a more direct relationship. In addition, their approach is heavily dependent on a single data source, where this paper presents a method of combining multiple data sources and classifiers in the subjectivity classification system.

(Hassan and Radev 2010) present a novel approach to randomly walking word relatedness graphs. The polarity score of each word is determined from the average time spent traversing the graph until a labeled instance is found. The polarity score mechanism presented in this paper is similar in nature because it represents the magnitude of positivity or negativity.

(Mandala, Tokunaga, and Tanaka 1999) inspired the idea for traversing WordNet. Their research aimed at improving information retrieval by combining WordNet synset words with thesauri information. They utilized the SMART stop word list (Lewis 2005), of which the stop words used by this paper are a subset (Switzer 2010).

The secondary focus of this research is the work of (Abedinm, Ng, and Khah 2010) towards causal analysis of the Aviation Safety Reporting System. They proposed using the Basilisk framework and a set of shaping factors to locate specific causes of incidents within each report of the corpus. Abedin provided the preprocessed ASRS corpus and the ASRS shaping factors training data.

3. Jargon Heavy Corpora

Domain specific languages present a different set of problems that would not otherwise be the present with more traditional corpora, such as the New York Times Annotated Corpus (Sandhaus 2008). The ASRS corpus uses a jargon heavy language that makes traditional natural language processing difficult. The Natural Language Processing community has not explored usage of jargon heavy corpus extensively. This paper will demonstrate that even though the corpus has a large number of words undefined by most English reference sources, we are able to utilize such a corpus for learning models.

(Posse et al. 2005) defined a set of shaping factors, a set of influential factors based on contextual information that will serve as our causal analysis classification labels. (Abedin, Ng, and Khan 2010). This paper aims to show that the reports can be classified a shaping factors based on the trained subjectivity lexicon.

The initial problem with the ASRS reports is the formatting of the report itself. Each report represents a dictated event describing a problem that occurred before, during, or after a flight. Since the reports are submitted electronically or retyped from mailed submissions, punctuation is rarely correct, words are often misspelled, sentence fragments are common, and misuse of capitalization is frequent. In comparison, the New York Times Annotated Corpus, created from New York Times publications, is properly formed, well written, and a highly regarded publication.

Another key obstacle to causal analysis of the ASRS corpus is the frequent usage of jargon words, or words that not typically found in dictionaries or thesauri. For example, the phrases “MECHS” and “nosewheel” are technical and specialized words not found by any of the reference dictionaries. The following excerpt demonstrates usage of jargon terms taken directly from the ASRS corpus.

> *Immediately descended below the lower layer to circling minimums instead of waiting to the outer marker.*
>
> *Received the Automatic Terminal Information Service at IAD with the following Notice To Airmen; taxiway whiskey closed between whiskey one and whiskey ten.*

Since most of the vernacular used in the reports relates to actions taken by a pilot while piloting a plane, there is a large number of words that have special meaning in context. Many of these terms refer to jargon not frequently seen outside of the context of aviation. From the quoted reported above, we can see the terms “circling minimums”, “outer marker”, “IAD”, “whiskey one” and “whiskey ten” all have special meaning to aviators (pilots and air traffic controllers).

This paper will present a technique that is a simplistic and effective means of automatic classification of ASRS shaping factors. While (Abedin, Ng, and Khan 2010) provided an equally effective means of classifying the most frequently associated words per shaping factor, their method of doing so was significantly more complex, relying on classifiers such as SVM.

This paper will demonstrate that identifying jargon words is unnecessary to directly process jargon. Instead, the method presented here focuses on the subjectivity of surrounding words.

This paper will also demonstrate that this method of the automatic classification can produce results on par with the human annotated methods. This is a significant and important step necessary to processing the high volume of yearly ASRS reports.

4. Subjectivity Classification

The first phase to processing jargon heavy corpora presented in this paper is to build a large and accurate subjectivity lexicon. We will focus on the subset of **polarity classification**, which is a measure of positivity or negativity, when performing subjectivity classification.

We shall subsequently use this lexicon to apply ASRS shaping factors (class labels). The automatic classification method presented here will be shown to discover phrases that frequently indicate shaping factors that human annotators may have dismissed.

As shown in Figure 1, the subjectivity classification system can be visualized as a directed graph. Each node represents a classification method or data source. Solid dark nodes represent the starting (or seeding) nodes and the directed edges represent the paths a subjectivity lexicon may travel during classification. To determine the set of experiments performed for this paper, start at each seeding node, the components based on overtly labeled data such as MPQA and ASL, and perform an acyclic traversal of the graph.

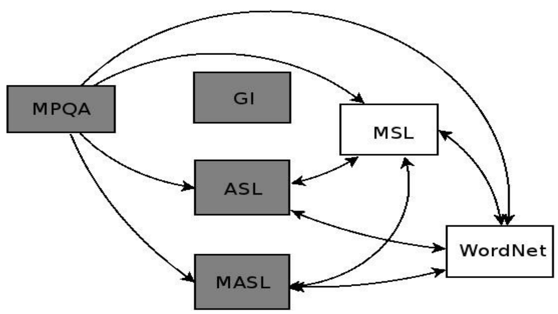

Figure 1. Subjectivity Classifiers

The General Inquirer, GI, lexicon will serve as the reference lexicon (except when boosting). For this reason, GI is not directly connected to any of the other nodes in Figure 1, and thus does not influence the subjectivity score of any other classifier.

The Multi-Perspective Question Answering, MPQA, lexicon is a set of polarity labels taken from GI and manually expanded with English reference sources. The MPQA data is overly labeled as being positive, negative, or neutral. The final polarity score is a summation of all of the identified polarity scores. This technique is a means of taking the consensus for words with different parts of speech and stemmed variations listed.

Another component is a dictionary-based method, known as the Affix Seed Lexicon, or ASL. This technique, based on the work of (Mohammad, Dorr, Dunne 2009), will look through a dictionary and find word pairs that match the 11 affix patterns. Since any word may be an unmarked positive word that pairs with a marked negative word, the runtime for this algorithm is $O(nm)$, where m is the number of affix patterns and n is the number of words in the dictionary data source.

A slight variation on the ASL algorithm is the Moby (wordlist) Affix Seed Lexicon, referred to as MASL. The purpose of this variation is to demonstrate the sensitivity of the ASL technique. The Moby Project is a public domain thesaurus and wordlists. The Moby wordlist is nearly 6 times larger than the unique words list taken from the Moby Thesaurus. Though this data source is much larger, we will show the performance is significantly lower, proving the ASL technique highly sensitive to noise.

Based on the Macquarie Semantic Orientation Lexicon, the Moby Seed Lexicon, MSL, algorithm follows the work of (Mohammad, Dunne, and Dorr 2009). We replaced the Macquarie Thesaurus with the Moby Thesaurus since the Macquarie Thesaurus is a commercial (non-free) data source. The entries in the thesaurus are treated as word clusters where the majority label is applied to the entire cluster. Equation 1 demonstrates how to compute the polarity scores. Note that new words are assumed neutral.

$$\delta = \sum_{\mathrm{j} \in \mathrm{W_i}} sign(polarity_i) \qquad (1)$$
$$\omega \in W_i : polarity_i = polarity_i + \delta$$

where δ represents the total change in the sign of the polarity score, ω is a word belonging to the set of root words W_i of word i, and $polarity_i$ is the polarity of word i.

The WordNet subjectivity classifier is similar to the thesaurus based learning method. By traversing through the synsets in WordNet, this algorithm can discover new words and label them based on prior labels.

The WordNet component works by iterating over all of the words and phrases in the prior subjectivity lexicon, and for each word, this algorithm will traverse WordNet in a depth first search (DFS) to find other related words. The set of synonyms taken as the related word cluster come from the "syns" sense key. The resulting set of related words, after searching to a depth of d, is conceptually the same as the manually clustered phrases from a thesaurus. The root word in this case is the word that starts the depth first search.

The polarity score for entire cluster of related words, regardless to relatedness depth, is modified using the same method as MSL. That is, the polarity score update rule in Equation 1 is applied to the entire WordNet relatedness cluster.

4.1. Boosting

Many of the component combinations perform poorly. For this reason, boosting provides a means of improving performance of these weak classifiers. In fact, that is the exact purpose of boosting algorithms, such as AdaBoost. A flexible boosting scheme is presented to further improve the accuracy without sacrificing lexicon size.

Similar to the initialization step of AdaBoost, we want to initialize the weights proportional to the size of the new additions. The new words will have equal initial weights proportional to the size of the new lexicon and prior words will be have their weights recomputed to match their proportion to the new lexicon. Equation 2 demonstrates the formula for recalculating the weights.

$$w(x_i) = \begin{cases} w_{m-1} * \dfrac{|x_\rho|}{|x_\varphi|}, & x_i \text{ is preexisting} \\ \dfrac{|x_\Delta|}{|x_\Delta| * |x_\varphi|}, & x_i \text{ is new} \end{cases} \quad (2)$$

where x_i is the lexicon instance, w_{m-1} is the prior weight for instance x_i, x_ρ is the prior lexicon, x_φ is the lexicon after classification, and x_Δ is the newly discovered lexicon phrases. When the size of the lexicon does not change, weight accommodation is unnecessary.

Directly applying the weight assigned by the AdaBoost algorithm to the polarity score has the opposite desired effect: misclassified instances will have higher polarity scores because AdaBoost increases the weight (for misclassifications). Equation 3 demonstrates a modification to the AdaBoost update rule, called **InvBoost**, to allow for direct application of the weight to the polarity score. This weight update rule works by de-emphasizing misclassifications and emphasizing correct classifications.

$$w_i = w_i * e^{\alpha_m * I(y_i \neq h_i(x_i))}, i = 1, 2, \ldots N$$
$$I(y_i \neq h_i(x_i)) = \begin{cases} 1, & if\ y_i\ is\ correctly\ classified \\ -1, & if\ y_i\ is\ misclassified \end{cases} \quad (3)$$

For this paper, an appropriate M, the number of boosting iterations, was determined experimentally. An arbitrary number of iterations that demonstrates an improvement over the baseline method was chosen.

5. Results

Since the MPQA data source was partially built based on the GI lexicon, we must use an alternative lexicon during testing in conjunction with the GI lexicon. This lexicon will be referred to as the Subjectivity Classification Manual Annotations lexicon, or SCMA. It was created by sampling 500 words and phrases from a lexicon created during development. Each of the words or phrases was assigned a polarity score by a human annotator. The human annotator had access to the definition of the word in the event that the word is new to their personal vocabulary level.

After the first human annotator labeled all 500 words, an additional human annotated the lexicon a second time, ensuring there is a high level of agreement concerning the labels. (Artstein and Poesio 2008) developed this technique. The SCMA lexicon was determined to have 92% agreement between both annotators, ensuring that the labels are reliably accurate.

Lastly, the SCMA lexicon will serve as the testing data when performing the boosting techniques. Since the GI lexicon is much larger than the SCMA lexicon, and many of the subjectivity classification combinations involve the MPQA technique, the GI lexicon will serve as the training data during boosting. This also avoids bias during the training phase since the MPQA data set originates from the GI lexicon.

5.1. Subjectivity Classification Results

The first set of experiments involves building the lexicons for each possible combination shown in Figure 1. Each experiment represents a single path through the component graph; each node visited represents a classifier applied to the prior subjectivity lexicon in the order of visitation. Table 1 lists each experiment and its respective performance. Note that "WNd2" stands for WordNet with a depth search of d = 2. Note that a modified set of stop words, neutral words such as prepositions, conjunctions, and pronouns, were omitted during training.

Table 1. Pre-Boosting Results

Pipeline	Size	vs. GI	
		Shared	F-1
ASL	8189	1025	83.6
MASL	56696	2168	72.9
ASL + MSL	87528	3276	70.4
ASL + MSL + WNd2	107598	3239	71.2
ASL + WNd2	15118	1847	76.3
ASL + WNd2 + MSL	98792	3376	69.8
MPQA	6436	2863	98.1
MPQA + ASL	12720	2875	96.5
MPQA + ASL + MSL	94337	3198	90.9
MPQA + ASL + MSL + WNd2	113366	3359	90.8
MPQA + ASL + WNd2	22892	3196	94.0
MPQA + ASL + WNd2 + MSL	101113	3469	92.3
MPQA + MSL	90391	3215	90.9
MPQA + MSL + ASL	88765	3237	91.5
MPQA + MSL + ASL + WNd2	107490	3375	91.0
MPQA + MSL + WNd2	108711	3351	90.2
MPQA + MSL + WNd2 + ASL	109159	3387	90.3
MPQA + WNd2	14379	3145	95.7
MPQA + WNd2 + ASL	19813	3211	95.4
MPQA + WNd2 + ASL + MSL	98489	3478	93.1
MPQA + WNd2 + MSL	98805	3485	92.3
MPQA + WNd2 + MSL + ASL	97185	3462	93.0

For the sake of brevity, we only show the results compared against GI in Table 1. The comparison of boosting results in Table 2 demonstrates a comparison against the SCMA lexicon.

The ASL method was surprising effective against the GI lexicon. While the ASL algorithm was effective alone, it was also effective in combination with other methods. When used early in the pipeline, the ASL technique found a small number of words with decent performance. When used later in the pipeline, ASL discovers fewer words, lowers the performance, and in some cases, reduces the total lexicon size. For example, the MPQA + ASL + WNd2 experiment discovers approx. 7,500 additional total words (over MPQA + WNd2), approx. 50 GI words, and reduces performance against GI by approx. 2%. The research from (Mohammad, Dunne, and Dorr 2009) proved valuable, as most of the words classified by this method are accurate.

The MASL method, ASL used with the Moby word list files, discovers significantly more words than ASL at the cost of a large performance decrease (16% lower against GI).

The MSL method discovers more words than any other classifier, regardless of the stage of execution in the experiment. However, unless it is combined with a high quality classifier, its performance is less than desirable. For example, ASL + WNd2 + MSL experiment, when compared to the ASL + WNd2 experiment, shows approx. 83,000 words discovered though the performance 84.8% accuracy to 59.7% (versus GI). However, the MPQA + ASL + WNd2 + MSL experiment, compared to MPQA + ASL + WNd2 experiment, discovers approx. 80,000 and only reduces the performance by approx. 9%. The MSL algorithm has also proven valuable in certain circumstances.

The WordNet method has been shown to be successful. The WNd2 method typically discovers approx. 15,000 words and usually only reduces the performance by 1-5%. Although it may discover fewer words than the other methods, its ability to maintain the performance of the prior lexicon while increasing the output lexicon size is a positive result.

The experiments involving MPQA are by far the best performing combinations. These combinations perform well against both the GI and SCMA reference lexicons. For example, the MPQA + ASL + MSL + WNd2 experiment produces a lexicon of comparable size to ASL + MSL + WNd2, though the performance goes from 59.6% to 82.4% when evaluated against SCMA. In fact, every experiment that involves MPQA has improved performance over the experiments excluding MPQA. This method has proven to be valuable source of seeding data.

Table 2 shows the F-1 scores of the AdaBoost and InvBoost techniques. For the M=3 boosting results, the AdaBoost algorithm reduced the performance, in comparison to InvBoost, in every experiment involving a classifier that considered the boosting weights during classification. For example, InvBoost improves the MPQA + ASL + MSL + WNd2 experiment results from 73.5 to 79. In fact, nearly every experiment demonstrates a marked improvement in subjectivity classification with InvBoost over AdaBoost (when directly applying the weights).

Table 2. Boosting Comparisons

Pipeline	AdaBoost F-1	InvBoost F-1
ASL	88.0	88.0
ASL + MSL	60.9	68.5
ASL + MSL + WNd2	59.6	70.1
ASL + WNd2	72.4	75.9
ASL + WNd2 + MSL	62.1	68.4
MPQA	96.8	96.8
MPQA + ASL	89.6	89.6
MPQA + ASL + MSL	72.7	77.5
MPQA + ASL + MSL + WNd2	73.5	79.0
MPQA + ASL + WNd2	83.8	84.0
MPQA + ASL + WNd2 + MSL	74.6	80.3
MPQA + MSL	65.9	74.9
MPQA + MSL + ASL	70.3	75.8
MPQA + MSL + ASL + WNd2	73.5	77.4
MPQA + MSL + WNd2	68.4	76.6
MPQA + MSL + WNd2 + ALS	69.3	77.0
MPQA + WNd2	80.4	79.3
MPQA + WNd2 + ASL	83.1	82.9
MPQA + WNd2 + ASL + MSL	73.2	79.0
MPQA + WNd2 + MSL	67.4	75.3
MPQA + WNd2 + MSL + ASL	71.3	76.7

5.2. ASRS Causal Analysis Results

To accompany the ASRS corpus, the ASRS training data set will serve a valuable purpose during this phase. The ASRS training data consists of reports that were manually labeled with one or more shaping factors. This information was utilized in conjunction with the Basilisk framework by (Mohammad, Ng, Khan 2010). The results presented here demonstrate that a subjectivity lexicon is useful in building a model of similar quality.

For this paper, each word and shaping factor will be evaluated by its information gain. We are interested in the set of words with the highest information gain for each shaping factor as they are indicative features of the shaping factor.

The information gain equation is interpreted as follows: instead of positive / negative labels, we will consider whether a shaping factor or word is absent or present in a report. The *Gain(S, A)* function will be calculated for each word A and every shaping factor S.

For the sake of brevity, Table 3 is a sampling of the words identified for a few of the shaping factors (there are 14 in total).

The most important observation is many of the words mirror the set of words listed by (Abedin, Ng, and Khan 2010). This is an indication that the subjectivity classification system is successfully capable of automatically identifying a shaping factor word list similar to the manually identified seed words.

Table 3. Classified Shaping Factor

Shaping Factor	Top words, sorted by Information Gain
Familiarity	familiarity, unfamiliar, unfamiliar with, airport, familiar, new, new to, training,

Shaping Factor	Top words, sorted by Information Gain
	familiar with, keep track of,visuals, particular, gentleman, especially, greatly, omission, stuff, concentrating, experience, time
Illusion	illusion, lights, allegedly, bargaining, black hole, blinding, breach, demarcation, exclusively, intensified, leverage, live, live with it, NB, one shot, reflecting, trap, lighting, judge, wall
Physical Environment	physical, environment, turbulence, weather, visibility, moderate, overcast, the weather, rain, thunderstorm, conditions, ice, cloud, wind, scattered, low, deteriorated, severe, decision, lightning

6. Conclusions

The prototype system of independent subjectivity classifiers presented demonstrates that a subjectivity lexicon is a valuable resource for determining root cause analysis over a jargon heavy corpus. The prototype system employed a set of semi-supervised subjectivity classifiers to build a large and accurate lexicon. We explored the complexities surrounding the ASRS corpus and showed that the information gain statistic is an effective means of identifying a set of words most frequently associated to a shaping factor.

Future development will focus in two primary directions: classification of the ASRS corpus and better data sources and classifiers in the subjectivity classification system.

The WordNet subjectivity classifier is currently restricted to identifying similar words and marking the entire cluster with the majority. The algorithm could be modified to mark the synonyms as it currently does and alter the polarity score of the set of antonyms in the opposite direction. That is, if the cluster of similar words indicates a positive cluster, then the antonym cluster would indicate a negative cluster.

The words that trigger high information gain for a given shaping factor could be used as features for training an SVM classifier. If a word with a high information gain for a shaping factor is located within an ASRS report, the classifier could be trained to assign the shaping factor as a label to the report.

The ASRS is a large and complex corpus. Since there is a large amount of jargon, it can be difficult to identify key features for any given report. However, this paper has shown that it is still possible to mine the corpus for primary indicators of problems.

7. References

[1] Mohammad, S., Dunne, C., and Dorr, B. 2009. Generating High-Coverage Semantic Orientation Lexicons From Overtly Marked Words and a Thesaurus. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 599-608, Singapore.

[2] Esuli, A. and Sebastiani, F. 2006. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, 417-422.

[3] Hassan, A. and Radev, D. 2010. Identifying Text Polarity Using Random Walks. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 395-403, Uppsala, Sweden.

[4] Mandala, R., Tokunaga, T., and Tanaka, H. 1999. Complementing WordNet with Roget's and Corpus-based Thesauri for Information Retrieval. In *Proceedings of the ninth conference EACL*, 94-101.

[5] Abedin, M. A. U., Ng, V., and Khan, L. 2010. Cause Identification from Aviation Safety Incident Reports via Weakly Supervised Semantic Lexicon Construction. In *Journal of Artificial Intelligence Research 38*, 569-631.

[6] Lewis, D. D. RCV1-v2/LYRL2004: The LYRL2004 Distribution of the RCV1-v2 Text Categorization Test Collection (14-Oct-2005). http://www.jmlr.org/papers/volume5/lewis04a/lyrl2004_rcv1v2_README.htm.

[7] Switzer, J. ASRS Stop Words. http://www.github.com/s1n/thesis/tree/master/data/smartstop.txt (created 17 November 2010).

[8] Sandhaus, E. 2008. The New York Times Annotated Corpus. Linguistic Data Consortium, Philadelphia.

[9] Posse, C., Matzke, B., Anderson, C., A., Matzke, M., & Ferryman, T. 2005. Extracting information from narratives: An application to aviation safety reports. In *Proceedings of the 2005 IEEE Aerospace Converence*, 3678-3690.

[10] Artstein, R., and Poesio, M. 2008. Inter-coder agreement for computational linguistics. In *Computational Linguistics, 34*(4), 555-596.

[11] Reynard, W. D., Billings, C. E., Cheaney, E. S., and Hardy, R. 1986. The Development of the NASA Aviation Safety Reporting System. In *NASA Reference Publication 1114*. California: NASA Scientific and Technical Information Branch.

[12] General Inquirer. http://www.wjh.harvard.edu/~inquirer/ (accessed 22 August 2010).

[13] Wilson, Theresa, Janyce Wiebe, and Paul Hoffman. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In Proceedings of HLT-EMNLP-2005.

[14] Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer Series in Statistics. New York: Springer Science+Business Media, LLC.