Sublanguage Analysis Applied to Trouble Tickets

Elizabeth D. Liddy, Svetlana Symonenko, Steven Rowe

Center for Natural Language Processing
School of Information Studies
Syracuse University
{liddy, ssymonen, sarowe}@syr.edu

Abstract

A feasibility study was conducted to determine whether the sublanguage methodology of NLP could analyze and represent the vital information contained in trouble tickets' ungrammatical text and to explore various knowledge mining approaches to render the data contained in these documents accessible for analysis and prediction. Experiments showed that the linguistic characteristics of trouble tickets fit the sublanguage theoretical framework, thus enabling NLP systems to tap into the unrealized value of trouble ticket data.

Motivation

Many large organizations have systems that manage the process of receiving and processing reports of trouble with their products, their services, or if they are a system provider, with their systems. However, most organizations are currently unable to fully leverage the value of the data contained in these problem reports, as well as in their company's responses, in order to gain proactive, adaptive insights about their products, customer services, or field service operations. This problem plagues a wide range of industries, from utilities to automotive manufacturers to financial services.

Trouble tickets (aka field service reports or problem reports) are typically a combination of structured and unstructured reports. While the structured data – database fields – may seem easier to exploit because of their supposed predictability, the range of language used in the structured fields is still a problem. Even more of a challenge are the unstructured portions of the reports, such as complaints, comment fields, and "remarks" because of the freer expression used in these sections in order to accurately describe a situation. Further complications arise from the numerous creative abbreviations and spellings found when textual data is input by a large number of individuals.

For most organizations, the only "tool" available for exploiting the information in trouble tickets is ad-hoc keyword searches, which often perform poorly. As a result, the business analyst who is trying to get a handle on trends and trouble spots typically resorts to reading many more trouble tickets than is really necessary. Furthermore, the analysis of such retrieved sets of trouble tickets is likely to be inaccurate if NLP is not used to capture the multiple ways that a common issue or solution has been described, thus resulting in the under-counting of issues or solutions.

Furthermore, without the capability to knit facts together across trouble tickets, the only organizing themes for tickets are static, such as date or gross categorization attributes. As a result, this potentially valuable information is organized and analyzed only in ways previously anticipated. That is, there is no facility for "knowledge discovery", so business analysts can find only what they explicitly ask for as they do not have tools that can surface emerging trends or previously unsuspected relationships.

Overview

This paper describes exploratory work conducted at the Center for Natural Language Processing (www.cnlp.org) as a feasibility study for a large utility provider. The company was interested in developing a knowledge discovery system to analyze the data aggregated by its Emergency Control System (ECS) in the form of field service tickets. When a "problem" in the company's electric, gas or steam distribution system is reported to the company's Call Center, an operator opens a new ticket. A typical ticket would contain the original report of the problem and all field work taken to fix the problem, as well as related scheduling and referring actions. Each ticket combines structured and unstructured data. Structured fields are fed from other information systems or filled by the operator in a menu-driven mode. Unstructured information, or free-text, is entered by the operator as s/he receives it over the phone from a person reporting a problem or a field worker dispatched to fix it. Based on the initial information about the problem, an operator also assigns a ticket an Original Trouble Type. This Trouble Type can be changed over the life cycle of a ticket, as additional information may clarify what the Actual Trouble Type is.

The overall goal of the feasibility study was to determine whether NLP could deal with this un-grammatical text and then to explore various knowledge mining approaches to facilitate exploitation of the knowledge in the tickets.

Related Research

Initial analysis of sample data suggested that the goal could be effectively accomplished from the perspec-tive of sublanguage theory. Sublanguage theory states that texts produced within a community engaged in a specialized activity deal with a particular subject area; share a common vocabulary; exhibit common, often unconventional, habits of word and grammar usage; feature a high frequency of certain odd constructions; and make extensive use of special symbols and abbreviations (Grishman & Kittredge, 1986; Harris, 1991). Sublanguage theory has been successfully applied in such domains as biomedicine (Friedman et al., 2002; Liddy et al., 1993), software development (Etzkorn et al., 1999), legal, aviation, weather forecasting (Somers, 2003), and others.

Trouble tickets, even while important to commer-cial applications, still get limited coverage in the literature. The most comprehensive description of trouble tickets (Johnson, 1992) points out that, although their main focus is on product- or service-related issues and actions undertaken to fix them, tickets may serve overall work planning purposes. Also, tickets exhibit a special discourse structure, combining system-generated, structured data and free-text sections; a special lexicon full of acronyms, abbreviations and symbols; and consistent "bending" of grammar rules in favor of speed writing (Johnson, 1992; Marlow, 2004).

Our exploration of data mining approaches applicable to trouble tickets texts has been informed by research on statistical techniques for mining domain-specific terms and phrases (Church & Hanks, 1989; Dunning, 1994; Godby, 1994); as well as machine learning techniques for multi-label class-ification (Joachims, 2002; Yilmazel et al., 2005a).

Methodology

For the feasibility study, the client provided us with a dataset of 162,105 trouble tickets dating from 1995 to 2005. About 80% of the tickets were less than 1 KB and a few tickets exceeded 25 KB. Data pre-processing operations included: (i) stripping ticket ID and line number from each line in a ticket body; (ii) converting non-ASCII characters to ASCII; (iii) con-verting tickets to XML format; & (iv) tokenizing text strings. The tokenizer was adapted to cover the identified special features of the trouble tickets' vocabulary and grammar, identified in the course of manual analysis of a data subset and included odd punctuation usage, common misspellings, name variants, abbreviations, and domain-specific phrases.

To explore whether the data's linguistic charac-teristics did fit the sublanguage model, a sample of data was manually annotated and analyzed. The sample included 70 tickets from a subset of 400 randomly selected tickets and 3 of the 6 largest tickets, all created in 2000-2005.

The manual annotation process aimed to identify consistent linguistic patterns in the texts, and in particular: (i) domain-specific vocabulary (abbre-viations and acronyms, special terms and phrases); (ii) major discourse components (ticket sections); and (iii) important semantic components (people, organizations, locations, timestamps, equipment, important concepts, etc.).

Identifying Core Domain Lexicon

Building a sublanguage model began with identifying the core domain vocabulary. Such a lexicon includes acronyms for Trouble Types (*SMH* - smoking manhole); department names (*EDS* - Electric Distribution); directions and locations (*N/W/C* - North West Corner; *S/S/C* - South of the South Curb); special terms (*PACM* - Possible Asbestos Containing Material, *LSE* - Life Sustaining Equipment Cus-tomer). Other lexicon classes contain abbreviations or shorthand, such as *BSMNT* (basement), *BLDG* (building), *F/UP* (follow up); and fixed phrases, such as *NO LIGHTS, WHITE HAT* (a person assigned in charge of a major incident, also referred to as I.C. – Incident Commander).

Since no established lexicon of this sort was maintained in the company, defining and compiling such lists involved close interaction with domain experts. Though originally intended to support the development of the sublanguage grammar, these lists can become part of the corporate knowledge base and provide additional value as a reference or training source, especially, for new employees, to improve the consistency of the manual data input.

Modeling Ticket Discourse Structure

Manual review of the data revealed a consistent structure for trouble ticket discourse. Figure 1 shows a typical trouble ticket.

Ticket 1

```
|001| CONST MGMT REPORTS SPARKING WIRE IN MH N/S SPRING ST
|002| 55' E/O 12TH AVE (ON WALK) - CONTRACTORS ON LOCATION-SJ
|003| 06/08/00 23:16 MDEJKSMITH DISPATCHED                  BY 12345
|004| 06/08/00 23:17 MDEJKSMITH ARRIVED                     BY 12345
|005| 06/08/00 23:17 CREW PULLED OFF FOR OUTAGE.....JS
|006| 06/08/00 23:18 MDEJKSMITH UNFINISHED                  BY 12345
|007| 06/09/00 15:00 MDEJLSMITH DISPATCHED                  BY 12345
|008| 06/09/00 16:00 MDEJLSMITH ARRIVED                     BY 54321
|009| 06/09/00 18:20 MDEJLSMITH REPORTS CLEARED MULTIPLE B/O'S
|010| IN SB#46977 N/S SPRING ST 55'E/O 12TH AV  READY FOR C.A.I -
|011| 06/09/00 18:34 MDEJLSMITH COMPLETE                    BY 54321
|012| 06/09/00 18:34 REFERRED TO: CAI     EDSWBR  FYI       BY 54321
|013| 06/10/00 14:10 NO C.M. ACTION REQD.==================BY 54321
```

Figure 1. A sample trouble ticket

Referring to Figure 1, every ticket consists of several text blocks normally marked by a "signature" of the operator, whether in the form of a numeric ID (*12345*) or initials (*JS*). Also, almost every ticket opens with a *complaint* (lines *001-002*) describing the original report of a problem. Such descriptions typically contain such components as: person reporting the problem (*CONST MGMT*), time, short problem description, location, equipment affected (*MH* – manhole), operator's ID (*JS*), and so on. A section describing the *field work* often features the name of the assigned employee, additional information about the problem, actions needed and/or taken to fix the problem, related complications (e.g. limited access or parking restrictions), etc. Lexical choices would also be fairly limited. For example, "reporting" is usually described as REPORTS or CLAIMS, STATES, CALLED; "completing the job" is normally recorded as JOB COMPLETE or OFFICE COMPLETION.

The analysis revealed a typical structure for trouble tickets (Table 1), in which sections are distinct in their purpose and the nature of data they contain. Not every ticket includes all of the identified sections and, as will be shown later, the analysis of the structure itself may lead to informative observations.

Section Name	Data
Complaint	Original report about the problem Free-text (input by the operator)
Office Action	Scheduling actions Structured text (generated by filling out formatted screens)
Office Note	
Field Report	Field work Free-text
Job Referral	Referring actions Closing actions Structured text
Job Completion	
Job Cancelled	

Table 1. Sample discourse structure of a ticket.

In parallel with identifying typical sections of a trouble ticket discourse, a schema of recurring, less explicit semantic components, such as people, locations, problem, timestamp, equipment, urgency, etc., was developed and each ticket was annotated with these schema elements.

Figures below show the resulting annotation of sample ticket by ticket sections (Fig.2) and of a *complaint* section by semantic components (Fig.3). For example, *CONST MGMT* is tagged as *complaint_source*, *MH* — as *ECS_structure* (an internal term referring to a piece of equipment), *JS* — as *entry_person*. Both schemas were validated by the customer's domain experts.

```
<complaint>
    CONST MGMT REPORTS SPARKING WIRE IN MH N/S SPRING ST
    55' E/O 12TH AVE (ON WALK) - CONTRACTORS ON LOCATION-JS
</complaint>
<office_action> 06/08/00 23:16 MDEJKSMITH DISPATCHED  BY 12345
  </office_action>
<office_note>
    06/08/00 23:17 MDEJKSMITH ARRIVED                 BY 12345
    06/08/00 23:17 CREW PULLED OFF FOR OUTAGE.....SJ
    06/08/00 23:18 MDEJKSMITH UNFINISHED              BY 12345
</office_note> ....
<field_report>
    06/09/00 18:20 MDEJLSMITH REPORTS CLEARED MULTIPLE B/O'S
    IN SB#46977 N/S SPRING ST 55'E/O 12TH AV  READY FOR C.A.I -
</field_report>
<job_completion>
    06/09/00 18:34 MDEJLSMITH COMPLETE                BY 54321
  </job_completion>
<job_referral>
    06/09/00 18:34 REFERRED TO: CAI  EDSWBR  FYI      BY 54321
</job_referral>
```

Figure 2. Annotated ticket sections.

```
<complaint_source> CONST MGMT </complaint_source>
  REPORTS
  <problem> SPARKING WIRE IN
    <ECS_structure> MH </ECS_structure>
    <location> N/S SPRING ST 55' E/O 12TH AVE (ON WALK)
      </location >
          CONTRACTORS ON LOCATION
  </problem >
          <entry_person> JS </entry_person>
```

Figure 3. Semantic components, *complaint* section

Analysis of manually annotated tickets, supple-mented with contextual mining of the entire dataset for particular terms and phrases, constituted the basis for developing rules for automatic identification of ticket sections and selected semantic components. Within the limits of the feasibility study, patterns for the components *time, feeder, ECS_structure, hazard, urgent* were fully implemented and tested.

The rules were then applied to automatically anno-tate a subset of 400 tickets. System performance was assessed on 70 manually annotated tickets and 80 un-seen tickets. Results demonstrated high accuracy of automatic section identification with an error rate of only 1.4%. There was no significant difference between results on the annotated vs. unseen tickets, showing that the texts have sublanguage-like limited variability and, thus, can be effectively analyzed and represented by a sublanguage-driven set of rules.

Evaluation of component identification was limited on some components such as urgent and hazard, which occurred only a few times, however, results for more frequent components, such as *feeder* and *ECS_ structure* (Table 2) are quite encouraging, given these are based on just the initial patterns.

Component	Precision	Recall
feeder	98%	93%
ECS_structure	88%	88%

Table 2. Results of automatic identification of *feeder* and *ECS_structure* components

The entire corpus of 162,105 tickets was next automatically annotated for sections and selected components. The resulting dataset became the *input* for n-gram analysis and machine learning experi-ments, which provided further empirical support for the sublanguage approach.

Practical Implications of Sublanguage

Such promising results of automatic identification of ticket sections and semantic components within ticket sections can, to a significant extent, be explained by the relatively limited number and high consistency of the identified linguistic constructions, which enabled their successful translation into a set of logical rules. This supported our initial view of the ticket texts as meeting the definition of a sublanguage In turn, the sublanguage approach enables the system to recognize and express a number of implicit semantic relationships in texts.

Mining for Frequency-Based Patterns

This section describes various knowledge discovery approaches that we tested in the belief that they would meet company needs and fit the nature of the data; reports their results; and discusses implications for the theoretical framework applied.

Patterns in Tickets' Discourse Structure

For each ticket, a string of ticket sections was generated, using the automatically labeled sections, which, essentially, exemplified the ticket's discourse structure (Figure 4).

Ticket 1:
complaint office_action job_completion

Ticket 2:
complaint job_referral office-note-add-info job_referral office-note-add-info job_referral office_action office-note job_completion

Figure 4. Examples of ticket section strings

Various kinds of descriptive analyses can be run using these structures. For example, we discovered that 54% of tickets do not have a *field-report* section, which means that a good half of the problems reported do not require dispatching a crew to the site. Relating such "no field-work" tickets to their Trouble Types may identify Types that typically do not involve field work, which, in turn, may bear implications for the company's overall planning processes.

Another example: in some tickets, the *field-report* section occurs after the *job*_completion, which may indicate that the problem was not fully fixed the first time. Such tickets may be of particular interest to the company's analysts looking for "weak spots" or areas of recurring problems.

N-gram Analysis

Initial review of the data contributed to shaping our original hypothesis that tickets belonging to different Trouble Types feature distinct linguistic character-istics. This was further empirically supported by an n-gram analysis, where an n-gram is a sequence of n consecutive words in text. The analysis was run, using the *Log-Likelihood* algorithm from the NSP tool (Banerjee & Pedersen, 2003), on the entire dataset of 162,105 tickets, with a stoplist applied to filter out prepositions, articles, pronouns, and other closed class words.

Review of bi-grams generated for the dataset (Table 3) supported our initial hypothesis that the trouble tickets make high use of a special vocabulary.

260921	MDE OFFICE
163816	DOCS JOB
90951	OFFICE COMPLETION
90353	OFFICE ARRIVED
90352	OFFICE DISPATCHED
85802	OFFICE OFFICE
57781	ACT TRBL
57780	TRBL CHNGD
43099	EDSFYI FYI
31407	MI.ES.BT ES0012

Table 3. Top 10 bi-grams, entire dataset

Focusing on the n-grams found in particular ticket sections such as the two in Table 4 provided empir-ical support to some section patterns identified in the course of manual analysis (highlighted in bold).

Complaint:	Field-report:
CUST STATES	B/O S
MANH D O REPORTS	**REPORTS FOUND**
ASAP ETS	MANH D O REPORTS
ASST ASAP	JOB COMPLETE
MANHOLE COVER	**SMITH REPORTS**
NETWORK PROTECTOR	GAS 0%
ELEC LIC	**BROWN REPORTS**
FIRE DEPT	**FLUSH ORDERED**
DEPT REPORTS	PSC MADE
REQ ASST	CO 0PPM

Table 4. Top 10 bigrams, *complaint* and *field_report*

Further focusing bi-gram analysis on particular Trouble Types supported our initial hypothesis that tickets of different Trouble Types use type-specific vocabulary in

their free-text sections. For example (Table 5), *Water Leak, No Lights, and AC Burnout* tickets each feature unique bigram phrases.

complaint:

WL – Water Leak	**NL – No Lights**	**ACB – AC Burnout**
WATER LEAKING	FUSES CHECKED	B/O S
WATER LEAK	PART SUPPLIED	DUCT EDGE
ASST ASAP	NO LIGHTS	AC BURNOUT
WATER COMING	LIC #	NO PARKING
REQ ASST	- RMKS	ACCESS ANYTIME

field-report:

WL – Water Leak	**NL – No Lights**	**ACB – AC Burnout**
WATER LEAK	PSC MADE	B/O S CLEARED
SUMP PUMP	B/O S	3-500 2-4/0
SERVICE DUCT	# 6	NO PARKING
DYE TEST	BLD #	FLUSH ORDERED
WATER LEAKING	3 PHASES	DUCT EDGE

Table 5. Top 5 bi-grams: *complaint* and *field-report* sections, three Trouble Types.

Results of our experimentation with n-gram statistics show that, in addition to providing an empirical foundation for the classification experi-ments, the n-gram analysis is useful for the task of generating a domain-specific lexicon.

Classification Experiments

Initial review of the dataset revealed that about 18% of all tickets were assigned the Miscellaneous (MSE) Trouble Type, which means that the problem described in such a ticket remains unclassified from the company's point of view. A number of reasons may account for such a situation, and in particular: (i) there is no more specific Type for this problem, because the problem is fairly unique; (ii) there is no more specific Type for this problem, because the problem, though not unique, for some reason, is not yet accounted for by the corporate classification system (ECS Trouble Types); or (iii) a more specific Type is available, but, for some reason, was not assigned by the Call Center operator, either initially, when the ticket was opened, or later, in the course of the ticket's life cycle. We proposed to investigate if the knowledge of type-dependent linguistic patterns can help with assigning specific Types to some MSE tickets.

The experiments were conceived of as providing a solution to real-world scenario iii, above, in which an NLP-enabled system, based only on the information in the *complaint* section, suggests to an operator a list of potentially relevant Trouble Types. This is a multi-label classification task, where the system is trained on problem descriptions from the *complaint* section of tickets belonging to specific Trouble Types and then tested on tickets belonging either to these Types or to MSE. Experiments were run using the *Extended LibSVM* tool, an extension of *LibSVM*, (Chang & Lin, 2001), modified and successfully tested in another project of ours (Yilmazel et al, 2005a; Yilmazel et al, 2005b).

In Experiment 1, the classifier was trained and tested on non-overlapping sets of tickets of the five most frequent non-miscellaneous Trouble Types with a Training-to-Test ratio at 75:25. For each Type, Training and Test vectors were generated from the *complaint* sections of the tokenized dataset, with a stoplist applied. In each vector, tickets of a particular Type constituted Positive examples, and all other tickets were Negative. For example, for Smoking Manhole (SMH), the Training vector contained 7432 Positive and 17659 Negative examples, and the Test vector included 2477 Positive and 5885 Negative examples. For each Type, the share of Positive exam-ples in Training and Test sets was the same (Table 6).

Trouble Type	**Type Description**	**Train / Test, tickets**	**Positive examples**
SMH	Smoking Manhole	7432 / 2477	30%
WL	Water Leak	5924 / 1974	24%
NL	No Lights	4184 / 1395	17%
OA	Open Auto (Feeder)	3751/ 1250	15%
ACB	AC Main Burnout	3800/ 1266	15%

Table 6. Positive examples, Training and Test sets

The results of Experiment 1 (Table 7) demonstrate the high performance of the classifier for precision and recall for both classes. We attribute this, to a great extent, to the fairly stable and distinct language – a sublanguage – of the trouble tickets.

Trouble Type	**Precision P, %**	**Recall P, %**	**Precision N, %**	**Recall N, %**
ACB	97.4	96.0	99.3	99.5
NL	96.8	95.4	99.1	99.4
OA	99.4	98.7	99.8	99.9
SMH	96.8	96.4	98.6	98.7
WL	98.6	98.0	99.4	99.6

P – positive class ("target" Type) N – negative class (the rest)

Table 7. Results of Experiment 1

In Experiment 2, the classifier was trained on 70,854 tickets from 20 specific Trouble Types and tested on 7,420

MSE tickets. As no "gold standard" was available, the evaluation had to be done manually and so it was limited to two Trouble Types: SMH and WL. For each Type, 50 tickets which the system had assigned the highest probability score were validated with the domain expert. Of 50 tickets classified into SMH, 24 were confirmed as "correct", and of 50 tickets classified into WL, 34 were confirmed as "correct". Analysis of misclassified tickets identified possible reasons for the classifier's mistakes, such as a limited number of Trouble Types in the Training set. As we move beyond the feasibility stage, we confidently expect these results to improve.

Conclusions and Future Work

Our initial exploration of trouble tickets revealed strong sublanguage characteristics, such as: wide use of domain-specific terminology, abbreviations and phrases; odd grammar rules favoring shorthand; and special discourse structure reflective of the commun-icative purpose of the tickets. The identified linguistic patterns are sufficiently consistent across the data that they can be described algorithmically to support effective automated identification of ticket sections and semantic components. Also, application of pattern-driven rules brings together name and spelling variants, thus streamlining and expanding coverage of subsequent data analysis.

N-gram analysis provided empirical support for classification experiments, but also demonstrated potential for the task of generating a domain-specific lexicon. Initial experimentation with machine learn-ing algorithms shows that applying the sublanguage theoretical framework to the task of mining trouble ticket data appears to be a promising approach.

Furthermore, machine learning-based classification using the identified sublanguage components can provide an effective and valuable solution. Firstly, a system can suggest to a Call Center Operator a list of possibly relevant Trouble Types for a new ticket to assist the operator in quickly choosing the Type to assign. Secondly, it will contribute to reducing the number of unclassified (MSE) tickets, both when the ticket is created and later, if classification is done off-line on already completed tickets. As a result, some of the MSE tickets can be brought into the scope of data mining techniques that use Type information.

Our directions for future research include experimenting with other machine learning techniques, i.e., mining for associations, utilizing to a greater extent the sublanguage-driven knowledge not only of ticket sections, but also of important semantic components.

References

Attensity. 2003. *Improving Product Quality Using Technician Comments.* White Paper.

Banerjee, S. & Pedersen, T. 2003. *The Design, Implementation, & Use of the Ngram Statistics Package.* Fourth International Conference on Intelligent Text Processing and Computational Linguistics, Mexico City.

Chang , C.-C. & Lin, C.-J. 2001. *LIBSVM: a library for support vector machines* http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Church, K. W. & Hanks, P. 1989. *Word Association Norms, Mutual Information & Lexicography*. Proceedings of 27th Annual Meeting of the Association for Computational Linguistics, Vancouver, B.C..

Dunning, T. 1994. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics, 19*(1): 61-74.

Etzkorn, L.H., Davis, C.G. & Bowen, L.L. 1999. The Language of Comments in Computer Software: A Sublanguage of English. *Journal of Pragmatics, 33*(11): 1731-1756.

Friedman, C., Kraa, P., & Rzhetskya, A. 2002. Two Biomedical Sublanguages: a Description Based on the Theories of Zellig Harris. *Journal of Biomedical Informatics, 35*(4): 222-235.

Godby, C.J. 1994. Two Techniques for the Identification of Phrases in Full Text. *Annual Review of OCLC Research*.

Grishman, R. & Kittredge, R. I. (Eds.). 1986. *Analyzing Language in Restricted Domains: Sublanguage Description & Processing*: Lawrence Erlbaum Assoc.

Harris, Z. 1991. *A theory of language and information: a mathematical approach.* Oxford: Clarendon Press.

Joachims, T. 2002. *Learning to Classify Text using Support Vector Machines: Ph.D. Thesis*: Kluwer Academic Pub.

Johnson, D. 1992. *RFC 1297 - NOC Internal Integrated Trouble Ticket System Functional SpecificationWishlist* http://www.faqs.org/rfcs/rfc1297.html.

Liddy, E. D., Jorgensen, C. L., Sibert, E. E. & Yu, E. S. 1993. A Sublanguage Approach to Natural Language Processing for an Expert System. *Information Processing & Management, 29*(5): 633-645.

Marlow, D. W. 2004. *Investigating Technical Trouble Tickets: An Analysis of a Homely CMC Genre.* 37th Hawaii International Conference on System Sciences.

Provalis. 2005. *Application of Statistical Content Analysis Text Mining to Airline Safety Reports*. White Paper.

Somers, H. 2003. Sublanguage. In H. Somers (Ed.), *Computers and Translation: A translator's guide.*

Yilmazel, O., Symonenko, S., Balasubramanian, N. & Liddy, E. D. 2005a. *Leveraging One-Class SVM and Semantic Analysis to Detect Anomalous Content.* IEEE International Conference on Intelligence and Security Informatics (ISI/IEEE 2005), Atlanta, GA.

Yilmazel, O., Symonenko, S., Balasubramanian, N. & Liddy, E.D. 2005b. *Improved Document Representation for Classification Tasks for the Intelligence Community*. Proceedings of the AAAI Stanford Spring Symposium.