Illuminating Trouble Tickets with Sublanguage Theory

Svetlana Symonenko, Steven Rowe, Elizabeth D. Liddy

Center for Natural Language Processing

School Of Information Studies

Syracuse University

Syracuse, NY 13244

{ssymonen, sarowe, liddy}@syr.edu

Abstract

A study was conducted to explore the potential of Natural Language Processing (NLP)-based knowledge discovery approaches for the task of representing and exploiting the vital information contained in field service (trouble) tickets for a large utility provider. Analysis of a subset of tickets, guided by sublanguage theory, identified linguistic patterns, which were translated into rule-based algorithms for automatic identification of tickets' discourse structure. The subsequent data mining experiments showed promising results, suggesting that sublanguage is an effective framework for the task of discovering the historical and predictive value of trouble ticket data.

1 Introduction

Corporate information systems that manage customer reports of problems with products or services have become common nowadays. Yet, the vast amount of data accumulated by these systems remains underutilized for the purposes of gaining proactive, adaptive insights into companies' business operations.

Unsurprising, then, is an increased interest by organizations in knowledge mining approaches to master this information for quality assurance or Customer Relationship Management (CRM) purposes. Recent commercial developments include pattern-based extraction of important entities and relationships in the automotive domain (Attensity, 2003) and text mining applications in the aviation domain (Provalis, 2005).

This paper describes an exploratory feasibility study conducted for a large utility provider. The company was interested in knowledge discovery approaches applicable to 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 corporate Call Center, a new ticket is created. A typical ticket contains the original report of the problem and steps taken to fix it. An operator also assigns a ticket an Original Trouble Type, which can be changed later, as additional information clarifies the nature of the problem. The last Trouble Type assigned to a ticket becomes its Actual Trouble Type.

Each ticket combines structured and unstructured data. The structured portion comes from several internal corporate information systems. The unstructured portion is entered by the operator who receives information over the phone from a person reporting a problem or a field worker fixing it. This free text constitutes the main material for the analysis, currently limited to *known-item* search using keywords and a few patterns. The company management grew dissatisfied with such an approach as time-consuming and, likely, missing out on emergent threats and opportunities or discovering them too late. Furthermore, this approach lacks the ability to knit facts together *across* trouble tickets, except for grouping them by date or gross attributes, such as Trouble Types. The company management felt the need for a system, which, based on the semantic analysis of ticket texts, would not only identify items of interest at a more granular level, such as events, people, locations, dates, relationships, etc., but would also enable the discovery of *unanticipated* associations and trends.

The feasibility study aimed to determine whether NLP-based approaches could deal with

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 169–172,
New York, June 2006.

such homely, ungrammatical texts and then to explore various knowledge mining techniques that would meet the client's needs. Initial analysis of a sample of data suggested that the goal could be effectively accomplished by looking at the data from the perspective of sublanguage theory.

The novelty of our work is in combining symbolic NLP and statistical approaches, guided by sublanguage theory, which results in an effective methodology and solution for such data.

This paper describes analyses and experiments conducted and discusses the potential of the sublanguage approach for the task of tapping into the value of trouble ticket data.

2 Related Research

Sublanguage theory posits that texts produced within a certain discourse community exhibit shared, often unconventional, vocabulary and grammar (Grishman and Kittredge, 1986; Harris, 1991). Sublanguage theory has been successfully applied in biomedicine (Friedman et al., 2002; Liddy et al., 1993), software development (Etzkorn et al., 1999), weather forecasting (Somers, 2003), and other domains. Trouble 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 work has also been informed by the research on machine classification techniques (Joachims, 2002; Yilmazel et al., 2005).

3 Development of the sublanguage model

The client provided us with a dataset of 162,105 trouble tickets dating from 1995 to 2005. An important part of data preprocessing included tokenizing text strings. The tokenizer was adapted to fit the special features of the trouble tickets' vocabulary and grammar: odd punctuation; name variants; domain-specific terms, phrases, and abbreviations.

Development of a sublanguage model began with manual annotation and analysis of a sample of 73 tickets, supplemented with n-gram analysis and contextual mining for particular terms and phrases. The analysis aimed to identify consistent linguistic patterns: domain-specific vocabulary (abbreviations, special terms); major ticket sections; and semantic components (people, organizations, locations, events, important concepts).

The analysis resulted in compiling the core domain lexicon, which includes acronyms for Trouble Types (*SMH* - smoking manhole); departments (*EDS* - Electric Distribution); locations (*S/S/C* - South of the South Curb); special terms (*PACM* - Possible Asbestos Containing Material); abbreviations (*BSMNT* - basement, *F/UP* - follow up); and fixed phrases (*NO LIGHTS, WHITE HAT*). Originally, the lexicon was intended to support the development of the sublanguage grammar, but, since no such lexicon existed in the company, it can now enhance the corporate knowledge base.

Review of the data revealed a consistent structure for trouble ticket discourse. A typical ticket (Fig.1) consists of several text blocks ending with an operator's ID (*12345* or *JS*). A ticket usually opens with a *complaint* (lines *001-002*) that provides the original account of a problem and often contains: reporting entity (*CONST MGMT*), timestamp, short problem description, location. *Field work* (lines *009-010*) normally includes the name of the assigned employee, new information about the problem, steps needed or taken, complications, etc. Lexical choices are limited and section-specific; for instance, reporting a problem typically opens with REPORTS, CLAIMS, or CALLED.

```
|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

The resulting typical structure of a trouble ticket (Table 1) includes sections distinct in their content and data format.

Section Name	Data
Complaint	Original report about the problem, Free-text
Office Action	Scheduling actions, Structured text
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.

Analysis also identified recurring semantic components: people, locations, problem, timestamp, equipment, urgency, etc. The annotation of tickets by sections (Fig.2) and semantic components was validated with 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.

The analysis became the basis for developing logical rules for automatic identification of ticket sections and selected semantic components. Evaluation of system performance on 70 manually annotated and 80 unseen tickets demonstrated high accuracy in automatic section identification, with an error rate of only 1.4%, and no significant difference between results on the annotated vs. unseen tickets. Next, the automatic annotator was run on the entire corpus of 162,105 tickets. The annotated dataset was used in further experiments.

Identification of semantic components brings together variations in names and spellings under a single "normalized" term, thus streamlining and expanding coverage of subsequent data analysis. For example, strings UNSAFE LADDER, HAZ, (hazard) and PACM (Possible Asbestos Containing Material) are tagged and, thus, can be retrieved as *hazard* indicators. "Normalization" is also applied to name variants for streets and departments.

The primary value of the annotation is in effective extraction of structured information from these unstructured free texts. Such information can next be fed into a database and integrated with other data attributes for further analysis. This will significantly expand the range and the coverage of data analysis techniques, currently employed by the company.

The high accuracy in automatic identification of ticket sections and semantic components 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 also supported our initial view of the ticket texts as exhibiting sublanguage characteristics, such as: distinct shared common vocabulary and constructions; extensive use of special symbols and abbreviations; and consistent bending of grammar in favor of shorthand. The sublanguage approach thus enables the system to recognize effectively a number of implicit semantic relationships in texts.

4 Leveraging pattern-based approaches with statistical techniques

Next, we assessed the potential of some knowledge discovery approaches to meet company needs and fit the nature of the data.

4.1 Identifying Related Tickets

When several reports relate to the same or recurring trouble, or to multiple problems affecting the same area, a note is made in each ticket, e.g.:

RELATED TO THE 21 ON E38ST TICKET 9999

Each of these related tickets usually contains some aspects of the trouble (Figure 3), but current analytic approaches never brought them together to create a complete picture of the problem, which may provide for useful associations. Semantic component *related-ticket* is expressed through predictable linguistic patterns that can be used as linguistic clues for automatic grouping of related tickets for further analysis.

Ticket 1
..REPORTS FDR-26M49 **OPENED AUTO** @ 16:54..
OTHER TICKETS **RELATED TO THIS JOB**
========= TICKET 2 ========== TICKET 3 =

Ticket 2
.. CEILING IS IN VERY BAD CONDITION AND IN DANGER OFCOLLAPSE. …

Ticket 3
.. CONTRACTOR IS DOING FOUNDATION WATERPROOFINGWORK ...

Figure 3. Related tickets

4.2 Classification experiments

The analysis of Trouble Type distribution revealed, much to the company's surprise, that 18% of tick-

ets had the Miscellaneous (MSE) Type and, thus, remained out-of-scope for any analysis of associations between Trouble Types and semantic components that would reveal trends. A number of reasons may account for this, including uniqueness of a problem or human error. Review of a sample of MSE tickets showed that some of them should have a more specific Trouble Type. For example (Figure 4), both tickets, each initially assigned the MSE type, describe the WL problem, but only one ticket later receives this code.

Ticket 1 Original Code="**MSE**" Actual Code="**WL**" WATER LEAKING INTO TRANSFORMER BOX IN BASEMENT OF DORM; …

Ticket 2 Original Code ="**MSE**" Actual Code ="**MSE**" … WATER IS FLOWING INTO GRADING WHICH LEADS TO ELECTRICIAL VAULT.

Figure 4. *Complaint* sections, WL-problem

Results of n-gram analyses (Liddy et al., 2006), supported our hypothesis that different Trouble Types have distinct linguistic features. Next, we investigated if knowledge of these type-dependent linguistic patterns can help with assigning specific Types to MSE tickets. The task was conceptualized as a multi-label classification, where the system is trained on *complaint* sections of tickets belonging to specific Trouble Types and then tested on tickets belonging either to these Types or to the MSE Type. Experiments were run using the *Extended LibSVM* tool (Chang and Lin, 2001), modified for another project of ours (Yilmazel et al., 2005). Promising results of classification experiments, with precision and recall for known Trouble Types exceeding 95% (Liddy et al., 2006), can, to some extent, be attributed to the fairly stable and distinct language – a sublanguage – of the trouble tickets.

5 Conclusion and Future Work

Initial exploration of the Trouble Tickets revealed their 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 communicative purpose of the tickets. The identified linguistic patterns are sufficiently consistent across the data, so that they can be described algorithmically to support effective automated identification of ticket sections and semantic components.

Experimentation with classification algorithms shows that applying the sublanguage theoretical framework to the task of mining trouble ticket data appears to be a promising approach to the problem of reducing human error and, thus, expanding the scope of data amenable to data mining techniques that use Trouble Type information.

Our directions for future research include experimenting with other machine learning techniques, utilizing the newly-gained knowledge of the tickets' sublanguage grammar, as well as testing sublanguage analysis technology on other types of field service reports.

6 References

Improving Product Quality Using Technician Comments.2003. Attensity.

Chang, C.-C. and Lin, C.-J. 2001. *LIBSVM* http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Etzkorn, L. H., Davis, C. G., and 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., and Rzhetskya, A. 2002. Two Biomedical Sublanguages: a Description Based on the Theories of Zellig Harris. *Journal of Biomedical Informatics, 35*(4): 222-235.

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

Harris, Z. *A theory of language and information: a mathematical approach.* (1991).

Joachims, T. *Learning to Classify Text using Support Vector Machines: Ph.D. Thesis* (2002).

RFC 1297 - NOC Internal Integrated Trouble Ticket System Functional Specification Wishlist.1992. http://www.faqs.org/rfcs/rfc1297.html.

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

Liddy, E. D., Symonenko, S., and Rowe, S. 2006. *Sublanguage Analysis Applied to Trouble Tickets.* 19th International FLAIRS Conference.

Marlow, D. 2004. *Investigating Technical Trouble Tickets: An Analysis of a Homely CMC Genre.* HICSS'37.

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

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

Yilmazel, O., Symonenko, S., Balasubramanian, N., and Liddy, E. D. 2005. *Leveraging One-Class SVM and Semantic Analysis to Detect Anomalous Content.* ISI/IEEE'05, Atlanta, GA.