
Illuminating Trouble Tickets with Sublanguage Theory

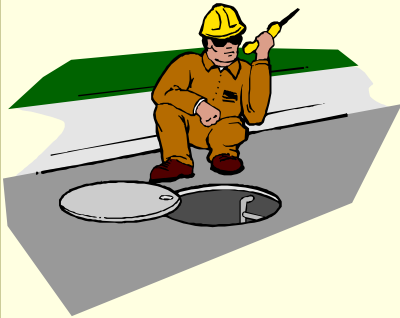
Svetlana Symonenko, Steven Rowe, Elizabeth D. Liddy

**Center for Natural Language Processing
School of Information Studies
Syracuse University**

June 6, 2006

Problem Space: Trouble Tickets

- Trouble tickets (help desk data) - reports to companies about problems with their products or services



ME00007923

|001| CONST MGMT REPORTS SPARKING WIRE IN MH N/S SPRING ST
|002| 55' E/O 12TH AVE (ON WALK) - CONTRACTORS ON LOCATION-MC
|003| 06/08/00 23:16 MDERWILLIM DISPATCHED BY 48414



- Current analysis of accumulated ticket data:
 - keyword searches or database queries
 - “untapped” value of hidden trends & anomalies
 - => missed proactive business insights
- Wide range of industries – utilities, automotive manufacturers, financial services...

Project Background

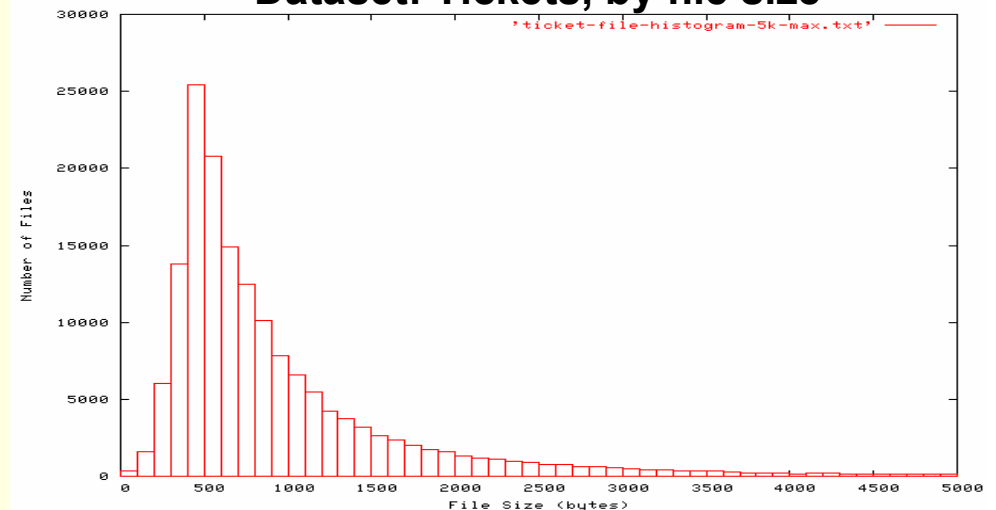
- Feasibility Study for a large utility provider
 - Trouble ticket - a “problem” in the company’s electric, gas or steam distribution system, reported to the Call Center
 - Current analysis – manual search for *known* patterns
 - Need to:
 - identify concepts of interest (events, people, locations...)
 - link the concepts across tickets to enable *knowledge discovery*



■ Dataset

- 162,105 tickets, from 2000-2005

Dataset: Tickets, by file size



Sublanguage-based Approach

- Sublanguage Theory - texts produced **within a community with a common purpose**, share:
 - specialized vocabulary
 - high frequency of odd constructions
 - deviant rules of grammar
 - predictable discourse structure

- Sublanguage-driven NLP analysis:
 1. Analyze DISTRIBUTION of words in a SAMPLE of tickets (73)
 2. Determine SEMANTIC WORD CLASSES
 3. Define sublanguage GRAMMAR based on words patterns
 4. Establish a SUBLANGUAGE MODEL based on identified lexicon, grammar & discourse structure.
 5. Specialize NLP TECHNOLOGY to accurately interpret texts

Common specialized vocabulary (1)

■ Abbreviations

- Trouble Types – *SMH* (Smoking Manhole)
 - Departments – *EDS* (Electric Distribution Service)
 - Directions, locations - *N/W/C* (Northwest corner)
 - Context-dependent:
 - *S/S/C* – Department (Subsurface Structure and Construction) or Direction (South of the South Curb)
 - *CO* – Company (*CO FORCES*) or Gas (*CO READINGS*)
 - Terms – *B/O* (Burn out), *C/F/R* (Cut For Replacement)
-

Common specialized vocabulary (2)

- special terms & phrases

- *FEEDER*



or



?

- *WHITE HAT*



or



?

*person in charge of a major incident...
also referred to as the I.C. or
Incident Commander.*

Domain-specific concepts

- Trouble Types (> 200):
 - *WL* (Water Leak)
 - *SMH* (Smoking Manhole)
- Locations:
 - C/P/W*
 - C/P WEST*
 - CENTRAL PK WEST*
 - CENTRAL PW*
 - CENTRAL PK W*
 - CENT PK WEST*
 - CENTRALPARK WEST*
 - CENTRAL PARK W*
 - = *CENTRAL PARK WEST*
- People: *UGSMITH* = *Smith* from the *UG* department)

Odd grammar

- Typos:

GREENWHICH = GREEWICH = GREENWITCH = GREENWICH

- Ellypsis of punctuation, subject, auxiliary verbs:

13:40HRS. ... CREW REQUESTED _ ==> TD =

*13:40HRS. ... CREW **IS** REQUESTED*

- Odd constructions:

*..WATER LEAKIN **N** BSEMENT =*

*WATER LEAKING **IN THE** BSEMENT*

*STAT CHG FR = **STATUS CHANGED FROM***

*... SUPER MADE TEST AND **ONE LEG DEAD**...*

Predictable structure: Ticket sections

“Raw” ticket

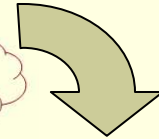
ME00007923

```

|001| CONST MGMT REPORTS SPARKING WIRE IN MH N/S SPRING ST
|002| 55' E/O 12TH AVE (ON WALK) - CONTRACTORS ON LOCATION-MC
|003| 06/08/00 23:16 MDERWILLIM DISPATCHED BY 48414
|004| 06/08/00 23:17 MDERWILLIM ARRIVED BY 48414
|005| 06/08/00 23:17 CREW PULLED OFF FOR OUTAGE.....CG
|006| 06/08/00 23:18 MDERWILLIM UNFINISHED BY 48414
|007| 06/09/00 15:00 MDEDONOHUE DISPATCHED BY 48414
|008| 06/09/00 16:00 MDEDONOHUE ARRIVED BY 44729
|009| 06/09/00 18:20 MDEDONOHUE 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 MDEDONOHUE COMPLETE BY 44729
|012| 06/09/00 18:34 REFERRED TO: CAI EDSWBR FYI BY 44729
|013| 06/10/00 14:10 NO C.M. ACTION REQD.===== BY 44979
    
```



Manual analysis



Ticket Sections

Section Name	Data
Complaint (<i>Initial Remarks</i>)	Free-text
Office Action	Structured text (produced by filling out formatted screens)
Office Note	
Office Note – Additional Field Information (<i>Ongoing Remarks</i>)	Free-text
Field Report	
Job Referral	
Job Completion	Structured text (produced by filling out formatted screens)
Job Cancelled	

Sample patterns: Complaint

- takes the 1st 1+ lines of a ticket;
- goes up to the next paragraph that:
 - begins with: Date? Time
 - ends with: BY (EmployeeID) OR 2-3 letters (initials)
- Pattern-based rules for automated identification of ticket structure
 - < 1.5 % error rate on 73 annotated and 80 unseen tickets

Predictable structure: Section components

Complaint

```
<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</location>
  CONTRACTORS ON LOCATION
</problem >
<entry_person> MC </entry_person>
```

- patterns for automated identification of some components: Time, Location, Hazard, Urgency..
- 90% accuracy on 70 annotated tickets
- conceptual grouping of variant expressions:

hazard:

05/07/04 08:01<INFO type="hazard"> **UNSAFE LADDER** </INFO>

WIRES /CABLES ..IS RUNNING ALONG THE SIDE WALK..CREATING A
<INFO type="hazard"> **HAZ** </INFO> ..WALKING

07/30/04 16:04<INFO type="hazard"> **PACM** </INFO> NEAR GAS SERVICE..

hazard:

(?<!NO) HAZARD(OUS)
HAZ
ASB
ASBESTOS
CO READING(S?)
CO\s*W*(?:=?)\s*W+PPM
BLOWN OFF
PACM

Knowledge Discovery: Mining Frequency-Based Patterns

- 2 and 3 grams for Trouble Types by particular ticket sections
 - Useful for mining domain-specific “lexico-syntactic phrases”
 - Trouble Types differ in their vocabulary in *complaint* and *field report* sections

Top 10 bi-grams - *Complaint* Section

WL	NL	ACB
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
ELEC CONDUIT	ENTIRE BLDG	CONST MGMT
ASAP ETS	CUSTOMER END	WEST WALL
CO ASST	ASST ASAP	EAST WALL
CHECK FIX	800-752-6633 BREAKERS	AC B/O
SWITCH GEAR	SUPPLIED ENTIRE	FLUSH REQUIRED

Knowledge Discovery: Related Tickets

- Refer to the same or recurring problem
- Each ticket has a “reference note” made in a predictable pattern:
 - Duplication: "(UN)?DUPLICATED TO (#)?(TicketID)"
 - Closure: "CLOSED (OUT)? TO (#)?(TicketID)"
 - Relation: "RELATED TO.. (#)?(TicketID)" , "SEE .. (#)?(TicketID)"
- Grouping of related tickets:
 - For a complete picture of the problem
 - For focused analysis

AZ01014381

MANH-DO REPORTS FDR-26M49 OPENED AUTO @ 16:54HRS 08/30/01...
08/31/01 01:30 HEYMACH O/G/S-FOB REPORTS FAULT ON FDR-26M49
IS A SECTION 1C-CABLE BETWEEN MH-6403 S/S W.34TH ST 130' W/O
5TH AVE. TO MH-6402 S/S W.34TH ST.290' W/O 5TH AVE.....KPM

OTHER TICKETS RELATED TO THIS JOB

===== AZ01014390 ===== AZ01014425 =====

AZ01014390

Z.SMITH OF FLUSH DEPT REPORTS AT S/S W.34ST 130' W/O 5AVE
MH-6403 **CEILING IS IN VERY BAD CONDITION AND IN DANGER OF
COLLAPSE.** 26M49 NEEDS TO BE TAGGED IN MH-6043 A.S.A.P.....AA

AZ01014425

WESSON REPORTS THAT IN OPEN TRENCH ON S/S W34ST 250'W/O
5AVE SEE FAULT ON FDR26M49 WHICH OPEN AUTO 08/30/01 16:54
**TICKET #AZ01014381 BELIEVE DAMAGE WAS DONE BY CHIPPING GUN
NO INJURIES REPORTED ... DID NOT OBSERVE DAMAGE ...
NOTE : THIS CONTRACTOR IS DOING FOUNDATION WATERPROOFING
WORK FOR THE EMPIRE STATE BLDG. THE ENTIRE PARKIN LANE IS
RIPPED OUT F/O EMPIRE STATE BLDG. EXPOSING OUR FACILITIES..**

Knowledge Discovery:

Machine Learning to assign *Trouble Type*

- Miscellaneous (MSE) is the top frequent Trouble Type:
 - 18% of tickets – out of scope for knowledge discovery
- In many cases, more specific *Trouble Types* could be assigned:

TICKET1 Original Code ="MSE" Actual Code ="WL"
WATER LEAKING INTO TRANSFORMER BOX IN BASEMENT OF DORM; PLS
CHECK FOR SAFETY

TICKET 2 Original Code ="MSE" Actual Code ="MSE"
... WATER IS FLOWING INTO GRADING WHICH LEADS TO ELECTRICAL
VAULT

- Machine Learning (classification):
 - Multi-label classification, Support Vector Machine (SVM)
 - System is trained on *Complaint* sections for known *Trouble Types*
 - For a new ticket, the system suggests the top-ranked Trouble Type
 - Can be done on-the-fly (for new tickets) or offline (for MSE tickets)

Classification Experiments (1)

- Dataset:

- Classifier is trained & tested on known Trouble Types (5 most frequent)

TT	Training	Test
EDSSMH	7432	2477
EDSWL	5924	1974
EDSNL	4184	1395
EDSOA	3751	1250
EDSACB	3800	1266

- Results:

Trouble Type	Precision	Recall
SMH	91.8	90.8
WL	98.4	97.9
NL	92.8	93.8
OA	99.7	98.7
ACB	93.2	88.6

Classification Experiments (2)

Dataset:

- Training- 21 known Trouble Types
- Test- MSE Type only

Evaluation:

- No “gold standard”
- Selective evaluation, for:
 - 50 tickets assigned “SMH”
 - 50 tickets assigned “WL”
- By analyst, confirmed with SME

Results:

- SMH: 24 “correct”
- WL: 34 “correct”

	Training	Test
EDSHCE	8247	0
EDSSMH	7432	0
EDSWL	5924	0
EDSNL	4184	0
EDSHME	3932	0
EDSUDC	3912	0
EDSACB	3800	0
EDSOA	3751	0
EDSSO	3684	0
EDSOPN	3036	0
EDSFLT	2621	0
EDSUAC	2612	0
EDSSOP	2545	0
EDSSLT	2409	0
EDSOOE	2300	0
EDSNLA	2291	0
EDSLV	2268	0
EDSTRF	2050	0
EDSSPD	2014	0
EDSWBR	1842	0
EDSMSE		7420

Utility of Sublanguage Approach

- Tickets' linguistic patterns are consistent
 - Can support automated identification of important ticket components (events, organizations, equipment..)
- Annotated data can be analyzed more effectively:
 - Section or component- focused mining for patterns
 - Use annotated data as input features to statistical analyses packages
 - Automatic assignment of specific Trouble Codes