

Sublanguage Analysis Applied to Trouble Tickets

Elizabeth D. Liddy, Svetlana Symonenko, Steven Rowe

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

May 11, 2006

Trouble Ticket Overview

- Field reports of problems customers encounter with a company's products, services, or systems
 - Trouble tickets / problem reports / help-desk data
 - Phone, in-person, company service person
- But companies are not fully leveraging the value of the data contained in these reports, including their company's responses
 - Unable to analyze & learn from trends that are only obvious when large repositories can be accurately represented for rich data mining
 - Miss out on proactive insights vital to their business interests
- Problem plagues a wide range of industries – from utilities, to automotive manufacturers, to financial services

Trouble Tickets

- Trouble Tickets are typically a combination of structured and unstructured sections
 - Structured portions - the range of language used is still quite varied
 - Unstructured portions of the reports – complaints, comment fields, remarks – exhibit even freer natural expression
 - Wide-ranging variety in expression due to multiple inputers
- **Business issue-** Current access to ticket contents is via ad hoc keyword searches or SQL style database queries
 - Routinely under-perform
 - System cannot deal with the noisy text & fails to capture multiple ways a common issue or solution occurs
 - Results in under-representation

Prototype System

- First customer is Con Edison
 - Utility provider to New York City & Westchester County
- Interested in developing a knowledge discovery system to analyze & learn from its aggregated field service tickets
 - Stored in Emergency Control System (ECS)
- Genesis of trouble tickets
 - A “Problem” in the company’s electric, gas, or steam distribution system is reported to the company’s Call Center
 - Operator opens a new trouble ticket
 - Field workers dispatched to fix it continuously input diagnosis & repair data
 - Base station may update with information coming from other sources

Prototype System (cont'd)

- Full trouble ticket contains:
 - Original report of the problem & all field work taken to fix the problem, as well as related scheduling & referring actions
 - Structured information from other information systems or menu-fills
 - Unstructured data 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 initial information about the problem, an operator also assigns a ticket an Original Trouble Type
 - Trouble Type can be changed over the life cycle of a ticket if additional information clarifies what kind of problem it is
 - Last Trouble Type assigned to a ticket becomes its Actual Trouble Type
 - Basis of class-level analysis

Sample Raw Trouble 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

Example: "Complaint"

COMPLAINT

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

Example: "Office Action"

OFFICE_ACTION

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

Example: "Office_Note"



OFFICE_NOTE

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

Example: "Field_Report"

FIELD_REPORT

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

Example: "Job_Completion"

JOB_COMPLETION

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

Example: "Job_Referral"

JOB_REFERRAL

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

Sublanguage Theory & Methodology

- Sublanguage Theory predicts that speech & texts produced within a community engaged in a specialized, common activity will:
 - Deal with a circumscribed subject area
 - Share a common vocabulary
 - Exhibit common habits of word usage
 - Use deviant rules of grammar
 - A subset of the rules of the standard language
 - High frequency of certain, odd constructions
 - Have a fairly predictable discourse-level structure
 - Make extensive use of special symbols & abbreviations
- Sublanguage Methodology
 - Specializes core NLP technology to application-specific requirements
 - Highly successful in developing models for multiple applications

Successful Sublanguage Applications

- Pharmacology reports
- Weather reports
- Technical manuals
- Cooking recipes
- Patents
- Stock market reports
- Patient medical histories
- Legal documents
- University catalogs
- Journal abstracts
- Life insurance applications
- Web pages

Prototypical Sublanguage Methodology

1. Select a representative SAMPLE of texts from a specialized community.
2. Conduct a DISTRIBUTIONAL ANALYSIS of words in sample texts.
3. Determine SEMANTIC WORD CLASSES
 - Reflecting the process / activities of the domain
 - Depend on words' similarities of occurrence.
4. Define sublanguage GRAMMAR based on co-occurrence patterns of sublanguage word classes.
5. Establish a SUBLANGUAGE MODEL based on sublanguage lexicon, grammar & discourse structure.
6. Specialize NLP TECHNOLOGY to accurately interpret new texts based on the Sublanguage Model.

Project Goals

- Goal of our project was to take this very messy data, semantically analyze it to normalize & tag the components of information reported in the trouble tickets
 - So all the data could then be statistically analyzed with assurance that all relevant data was included
 - Company could extract retrospective & predictive value from all its information assets
- Demonstrate that methodology was scalable & viable
 - With insight into extensibility to other organizations' Trouble Ticket problems

Phase 1A: Data Cleanup / Pre-processing

- Dataset provided- 162,105 tickets:
 - *Data* file (structured data, generated semi-automatically)
 - *Remarks* file (free-text data, entered by Call Center Operators)
 - From 2000 - 2005
- Pre processing steps:
 - Stripped Ticket ID and line # from each line
 - Converted non-ASCII characters
 - For each ticket, combined data components from *Remarks* file and *Data* file
 - Converted tickets to XML format
 - XML markup for ticket sections & section components

Phase 1B: Develop the Tokenizer

- Adapted our tokenizer for utility trouble ticket data:
 - To cover the special features of Con Edison language usage, such as name variants, common misspellings, abbreviations, fixed phrases, etc
 - Token = a term, including acronyms and fixed phrases, e.g.:
MANHOLE, NO ACCESS, I & A
 - 17 ways to abbreviate Broadway
 - Need to be adaptive to recognize unseen variants
- Sampled a subset of 400 randomly selected tickets + 6 largest (over 23Kb each) tickets created in 2000-2005
- Annotated & analyzed a sample of 70 tickets
 - From the 400 subset +3 from the set of 6 largest tickets)

Phase 2: Identify Explicit 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	Structured text (produced by filling out formatted screens)
Job Completion	
Job Cancelled	

Resulting Annotation: Ticket Sections

<complaint>

CONST MGMT REPORTS SPARKING WIRE IN MH N/S SPRING ST
55' E/O 12TH AVE (ON WALK) - CONTRACTORS ON LOCATION-MC

</complaint>

<office_action> 06/08/00 23:16 MDERWILLIM DISPATCHED BY 48414 </office_action>

<office_note>

06/08/00 23:17 MDERWILLIM ARRIVED BY 48414

06/08/00 23:17 CREW PULLED OFF FOR OUTAGE.....CG

06/08/00 23:18 MDERWILLIM UNFINISHED BY 48414

</office_note>

<office_action> 06/09/00 15:00 MDEDONOHUE DISPATCHED BY 48414 </office_action>

<office_note> 06/09/00 16:00 MDEDONOHUE ARRIVED BY 44729 </office_note>

<field_report>

06/09/00 18:20 MDEDONOHUE 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 MDEDONOHUE COMPLETE

BY 44729

</job_completion>

<job_referral>

06/09/00 18:34 REFERRED TO: CAI EDSWBR FYI

BY 44729

</job_referral>

<office_note> 06/10/00 14:10 NO C.M. ACTION REQD.===== BY 44979 </office_note>

Phase 3: Sublanguage Model Development

Lexical: Identify core vocabulary & deviant syntax:

- Abbreviations & acronyms
 - Trouble Types - *SMH*
 - Departments – *EDS, S/S/C*
 - Directions, locations - *N/W/C, S/S/C*
- Special terms & uncommon phrasings
 - *FEEDER, WHITE HAT, CO FRCES*

Semantics:

- Distributional lexical analysis to identify core domain concepts (Trouble Types; Location; Time; Person; ConEd Department, etc)

Discourse:

- Identify tickets discourse structure (major sections; section components)

Phase 4: Automatic Ticket Representation: Ticket Sections

- Developed & implemented rules for automatic identification of discourse structure of tickets
- Ran automatic ticket section identification on the entire dataset
- Manually evaluated a sample of the system output (73 annotated and 80 unseen tickets)
 - < 1.5 % error rate
- **Conclusion:** System can recognize predictable discourse level structure of tickets

Phase 4: Automatic Ticket Representation: Semantic Components

- Implemented patterns for recognition of 7 implicit semantic components

Time	Location	Entry_Person	Feeder
ECS_structure		Hazard	Urgency

- Compared system output to 70 'gold standard' manually tagged tickets
 - System accuracy shown to be 90% or higher
 - Varied based on particular component
 - Proving that automatic identification of semantic components can be done effectively
- Tagging at semantic component level brings together variant lexico syntactic expressions of same meaning

Semantic Components - Complaint Section

**CONST MGMT REPORTS SPARKING WIRE IN MH N/S SPRING ST
55' E/O 12TH AVE (ON WALK) CONTRACTORS ON LOCATION.
MC**

Semantic Components - Complaint Section

**CONST MGMT REPORTS SPARKING WIRE IN MH N/S SPRING ST
55' E/O 12TH AVE (ON WALK) CONTRACTORS ON LOCATION.
MC**

<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> MC </entry_person>

Phase 5: Applying Knowledge Discovery Approaches

- To gain a broader, more accurate picture of past & present utility problems in NYC, needed to:
 - Group related tickets
 - Mine for associations
- Could then analyze high frequency problems by:
 - Location
 - Time
 - Urgency level
 - Trouble Type
- Applied Machine Learning to:
 - Assist with Trouble Type assignment

Mining Frequency-Based Patterns: In Free Text *Remarks* File

- Mined for n-grams of consecutive terms using the Log-likelihood algorithm
- Bi-grams and tri-grams were generated for Trouble Types by particular ticket sections
- Analysis of bigrams & trigrams:
 - Useful for mining domain specific “lexico syntactic phrases” revealing common problems, locations, etc.
 - Showed that Trouble Types differ in their vocabulary in both *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

Machine Learning for Assigning *Trouble Type*

Machine Learning for Assigning *Trouble Type*

- MSE is the most frequent *Trouble Type* – 18%, which means:
 - Almost 1/5 of all tickets cannot be effectively mined for associations between a *Trouble Type* and other ticket components
- We demonstrated that in many cases, more specific *Trouble Types* could be assigned:

```
<TICKET id="ME05003448" original-code="EDSMSE" actual-code="EDSWL">  
<SECTION type="complaint"> WATER LEAKING INTO TRANSFORMER  
BOX IN  
BASEMENT OF DORM; PLS CHECK FOR  
SAFETY
```

Patterns in Trouble Code Changes

Top 10 *Original* Trouble Codes

29678	18.3%	EDSMSE
10996	6.8%	EDSHCE
9909	6.1%	EDSSMH
7898	4.9%	EDSWL
5579	3.4%	EDSNL
5242	3.2%	EDSHME
5216	3.2%	EDSUDC
5066	3.1%	EDSACB
5001	3.1%	EDSOA
4912	3.0%	EDSSO

Patterns in Trouble Code Changes

Top 10 Original Trouble Codes

29678	18.3%	EDSMSE
10996	6.8%	EDSHCE
9909	6.1%	EDSSMH
7898	4.9%	EDSWL
5579	3.4%	EDSNL
5242	3.2%	EDSHME
5216	3.2%	EDSUDC
5066	3.1%	EDSACB
5001	3.1%	EDSOA
4912	3.0%	EDSSO

Top 10 Actual Trouble Codes

29032	17.9%	EDSMSE
9111	5.6%	EDST9X
7928	4.9%	EDSWL
4780	2.9%	EDSOA
4691	2.9%	EDSACB
4378	2.7%	EDSNL
4288	2.6%	EDSUDC
4203	2.6%	EDSSMH
4014	2.5%	EDSOPN
3589	2.2%	EDSUAC

Patterns in Trouble Code Changes

Top 10 Original Trouble Codes

29678	18.3%	EDSMSE
10996	6.8%	EDSH C E
9909	6.1%	EDSSMH
7898	4.9%	EDSWL
5579	3.4%	EDSNL
5242	3.2%	EDSHME
5216	3.2%	EDSUDC
5066	3.1%	EDSACB
5001	3.1%	EDSOA
4912	3.0%	EDSSO

Top 10 Actual Trouble Codes

29032	17.9%	EDSMSE
9111	5.6%	EDST9X
7928	4.9%	EDSWL
4780	2.9%	EDSOA
4691	2.9%	EDSACB
4378	2.7%	EDSNL
4288	2.6%	EDSUDC
4203	2.6%	EDSSMH
4014	2.5%	EDSOPN
3589	2.2%	EDSUAC

Learning Patterns for Trouble Type Assignment

- Using Machine Learning:
 - Trained a system on 'problem descriptions' for known classes of *Trouble Types*
- For a new ticket, the system either:
 - Assigns the top-ranked *Trouble Type*, based on its NLP-based learning of 'problem descriptions'
 - Or can suggest list of possible *Trouble Types* for operator to select from

Learning Patterns for Trouble Type Assignment

- Using Machine Learning:
 - Trained a system on ‘problem descriptions’ for known classes of *Trouble Types*
- For a new ticket, the system either:
 - Assigns the top-ranked *Trouble Type*, based on its NLP-based learning of ‘problem descriptions’
 - Or can suggest list of possible *Trouble Types* for operator to select from
- Can also be done “off-line” for pre-existing tickets to clear up the backlog of MSE tickets and increase the number of tickets with real *Trouble Types* for data mining

Learning Patterns for Trouble Type Assignment

- Using Machine Learning:
 - Trained a system on ‘problem descriptions’ for known classes of *Trouble Types*
- For a new ticket, the system either:
 - Assigns the top-ranked *Trouble Type*, based on its NLP-based learning of ‘problem descriptions’
 - Or can suggest list of possible *Trouble Types* for operator to select from
- Can also be done “off-line” for pre-existing tickets to clear up the backlog of MSE tickets and increase the number of tickets with real *Trouble Types* for data mining
- Highly promising results on 6,500 unseen *Trouble Tickets*

Machine Learning Experiments

- Operational Scenario
 - An NLP-enabled system, based on information in the *Initial Remarks* ('*complaint*') section, suggests to a Call Center Operator a list of potentially relevant Trouble Types
- Algorithm: Support Vector Machine (*SVM*)
- Experimental design:
 - Multi-label classification task
 - System was trained on problem descriptions from '*complaint*' section of specific Trouble Type tickets
 - Experiments
 1. Classifier is tested on tickets with known Trouble Types
 2. Classifier is tested on miscellaneous tickets

Machine Learning: Experiment 1

- Train & test classifier on known Trouble Types
 - Using “gold standard” available from the client
 - 5 most frequent Original Trouble Types selected
 - ‘Complaint’ section used because it contains only the information available to the Call Center Operator at the time the ticket is created and assigned a Trouble Type

- *Dataset:*

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

Machine Learning: Experiment 1 Results for the “Target” Type

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

Machine Learning: Experiment 2

- Trained classifier on 5 known Trouble Types from Experiment 1
- Ran tickets that had been marked as MSE by Con Ed operators through the classifier
 - 7420 MSE tickets in the Test set
- Manually evaluated the results with validation by Con Ed's SMEs

Machine Learning: Experiment 2 Dataset

	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

Machine Learning: Experiment 2 Results

- Of 7420 MSE tickets in the Test set, system classified:
 - 181 as SMH
 - 330 as WL
- For each Type, 50 tickets were manually evaluated
- Of 50 MSE tickets classified into SMH:
 - 25 were judged by CNLP analyst as “correct”
 - 14 of these had had their original MSE Type later changed to SMH
 - Of remaining 11 tickets validated with the SME, 10 were confirmed as SMH
- Of 50 MSE tickets classified into WL:
 - 35 were judged by CNLP analyst as “correct”
 - 23 of these had had their original MSE Type later changed to WL
 - Of remaining 12 tickets validated with the SME, 11 were confirmed as WL
- Results are highly promising
 - Improving consistently with each iteration of learning, as other types are added to the training set

Enabling Business Insights

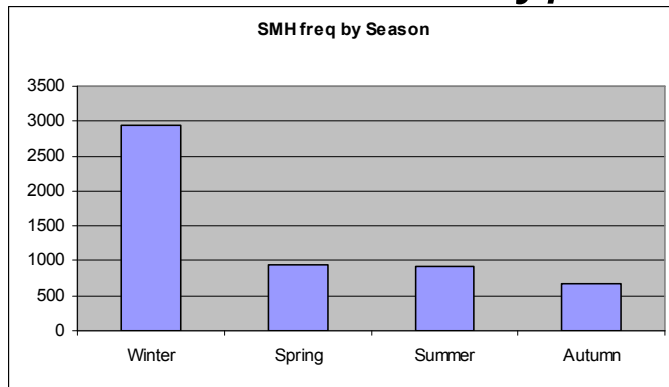
- Prior, the only organizing themes for tickets were static
 - Such as date or gross categorization attributes
 - Potentially valuable information could only be organized and analyzed in ways previously anticipated
 - Could find only what they explicitly asked for
 - Dramatic under-counting due to inability to recognize semantic commonalities
 - Data was, therefore, under-utilized

Enabling Business Insights

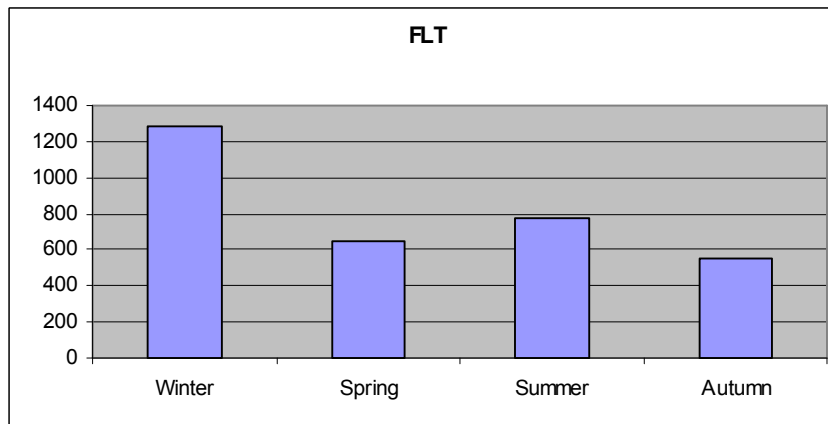
- Prior, the only organizing themes for tickets were static
 - Such as date or gross categorization attributes
 - Potentially valuable information could only be organized and analyzed in ways previously anticipated
 - Could find only what they explicitly asked for
 - Dramatic under-counting due to inability to recognize semantic commonalities
 - Data was, therefore, under-utilized
- Now, business analysts can get a handle on trends & trouble spots
 - Able to knit facts together across trouble tickets
 - Facility for “knowledge discovery”, as they use tools that can surface emerging trends or previously unsuspected relationships
 - Can apply full range of statistical analyses on ‘good’ data
 - Informed prediction now possible

Seasonal Patterns Discovered - 1

- Some *Trouble Types* show distinct seasonal patterns:

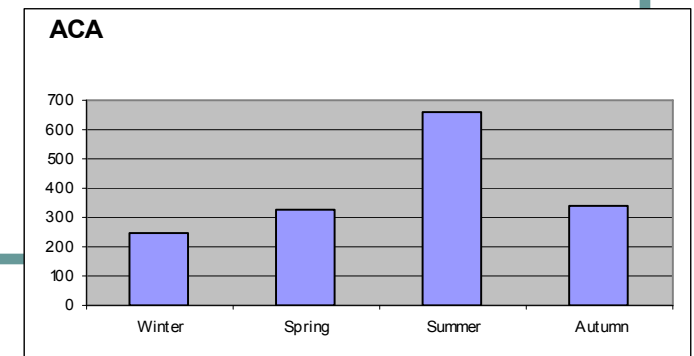


Smoking Manhole



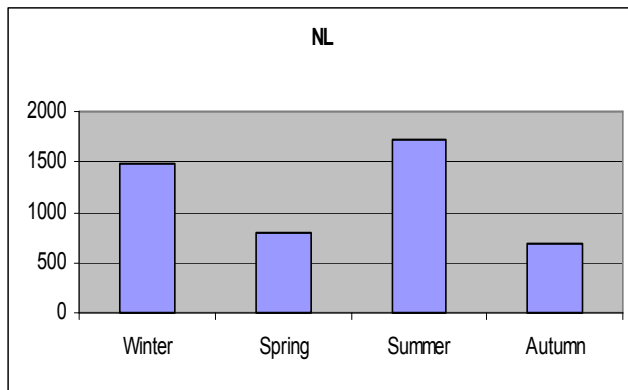
Flickering Lights

Asbestos Clear Access

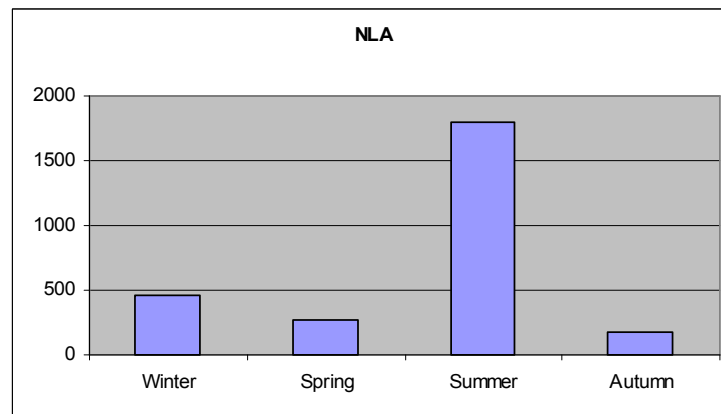


Seasonal Patterns Discovered - 2

- Noticeable differences in seasonal patterns of similar *Trouble Types*:



No Lights - Individual



No Lights - Area

High Frequency All-Season Locations

Frequency	Street / Ave		Cross Street / Ave
760	YORKVILLE		HELLGATE SUBSTATION
611	BROADWAY		
202	57	ST	5 AV
188	WATER	ST	COENTIES SLIP
162	1	AV	E 40 ST
158	125	ST	5 AV
155	47	ST	5 AV
151	BROADWAY		SPRING ST
150	57	ST	AMERICAS AV
148	54	ST	5 AV

Findings & Conclusions

- Language of Con Edison's Trouble Tickets demonstrates utility of sublanguage approach
 - Tickets' linguistic patterns are consistent & can be utilized to support semi- or fully-automated identification of important ticket components (events, organizations, equipment, urgency)
 - Bringing together lexical & syntactic variants streamlines and expands coverage of subsequent data analysis.
- Utilizing identified components & sublanguage patterns, can analyze data more effectively:
 - High-level performance in automatic assignment of specific Trouble Codes
 - Using annotated data as input to sophisticated statistical analyses packages
 - Temporal & seasonal data analysis may lead to insights into factors affecting Con Edison's proactive plans

Next Step - Indicators of 'Severity' of Case

- Trouble Type
- # of *field-persons* involved;
- Whether an additional crew is requested
- Whether external agencies and companies are involved
- Length of text:
 - Entire ticket
 - *Field-report* sections
- Duration (time span) of a case
- Whether the ticket has been re-opened
- Max # of clients interrupted
- # of calls received for the ticket
- Presence of certain clues (e.g. *Urgency* or *Hazard* section components)

THANKS!

- Questions?