DATA SCIENCE FOR SOCIAL GOOD

RAYMOND NG
DIRECTOR OF DATA SCIENCE INSTITUTE
PROFESSOR OF COMPUTER SCIENCE
UNIVERSITY OF BRITISH COLUMBIA
(AUGUST 5, 2019)

Outline for the Short Course

1. **Overview of the DSSG Program**
2. Theme 1: Transportation
3. Theme 2: Housing and Urban Development
4. Theme 3: Disease Control and Laboratory Testing
5. Key Technical Foundation: Natural Language Processing
6. Conclusions

DSSG Vision

Vision: To facilitate a quantum leap in society's ability to gain value from data by

- enhancing the capability of students to learn from data
- inspiring many more students towards further study of data science
- in turn dramatically increasing the flow of data-science-capable specialists into the workforce, **particularly working with not-for-profit organizations**

DSSG Mission: Student Teams working on Data-rich Projects that have societal benefits

- 4-month full-time data science capstone projects
- Interdisciplinary and immersive research experience combining technical data science training with social and ethical awareness
  - Diverse teams of graduate and undergraduate students from STEM and non-STEM fields
  - Hands-on data science training with open and private data sets
  - Awareness of various ethical and privacy issues around use of citizen data
- Projects with government or non-profit organizations

What is Data Science (DS)?

- Refers to methods, processes and tools that allow a user to extract useful information from complex – often large – data collections
- Is an interdisciplinary field involving many areas in computer science, statistics and mathematics:
  - Data mining, machine learning, statistical inference, predictive modeling, databases, visualization, high performance computing, linear algebra, data privacy, security, etc.
- Is critical for organizations and companies to make decisions and drive innovation

Why Data Science – the 4V's?

- Volume – rapid accumulation of data, partly because of the advances of devices collecting or generating data
- Variety – not just alpha-numeric data, but also text, images, time series, networks, etc.
- Velocity – the speed that data are collected or generated
- Veracity – the truthfulness/reliability/quality of data and data sources

Data Science Meets Smart Cities

Projects used DS tools for:

- Linking and Integrated modeling of diverse types of data, e.g., geo-spatial, text
- Extraction of sentiment and other related terms
- Cleansing and pre-processing of large amounts of data
- Compliance of individual privacy

- Diverse data now easily collectible by cheap devices, e.g., sensors, cameras
- Form the basis of transparent decision and policy making
- Find creative solutions to problems important to residents
- Maximize use of scarce resources and save costs

Program Overview

Week 1 — 2: Orientation, Workshops & Meet Project Sponsors

- Intermediate R/Python, data reproducibility, statistical modelling, machine learning
- Establish scope of work, deliverables, timeline

Week 3 — 12: Project Work

- Team-based research in consultation with Project Sponsors
- Regular meetings with Project Sponsors and DSI Scientific Director
- Invited Speaker Series (topics related to projects and methods)
- Mentoring by statistical consultants, industry data scientists/ developers, DSI postdocs and faculty

Week 13 — 14 : Conclusion

- Documentation (i.e., repository, report)
- Final presentation

Multiple Levels of Outside Sponsorship

- Sponsors in 2017 and 2018: Microsoft and Mitacs Accelerate (for graduate students)
- Sponsors in 2019: Boeing, Mitacs, UBC DSI
- In-kind contributions
  - sponsors offering career development advice
  - project partners offering data and staff time
  - alumni offering talks

Sponsors

BOEING

Microsoft

Mitacs

Project Partners

BC Centre for Disease Control

BRITISH COLUMBIA

CITY OF SURREY the future lives here.

Impact: Partners and Social Good

- Inform partners of their data collection processes (i.e., what data they need to capture and with what frequency)
- Tools developed being used as prototypes in the partner organizations, e.g., visualization tools, or pipelines to automate certain tasks
- Fellows hired by partners, such as the City of Surrey, BC CDC, etc.
  - Data scientists are expensive to hire
  - Governments and not-for-profit organizations find it hard to match salaries
  - Data scientists who are willing to take a "pay cut" need to have the right "mindset":
    - ***Use your skills (however temporarily) to help better the environment surrounding you***

Impact: Research and Curriculum

- In 2017 and 2018, DSSG teams gave poster presentations of their projects and results in the Cascadia Innovation Conference
- Two manuscripts from 2018 projects were submitted to the applied research track of the 2019 ACM SIGKDD conference
- The data pipeline developed by the Surrey rental housing project is being used by UBC School of Community and Regional Planning
- Datasets and tools can be used by undergraduate and graduate courses, e.g., urban studies, geography and CS

Multiple Levels of Mentoring for Fellows

- I met with each project team individually at least twice a month
- Postdoctoral fellows provided their expertise in selected areas of needs, e.g., NLP, ML, record de-duplication
- Industry mentors (e.g., Microsoft, Boeing) met with each team about once every four weeks and provide technical and non-technical mentoring
- Project sponsors met with their teams once a week about domain knowledge and business needs

Conclusions: Overview

- Very popular for students, project partners and sponsors
- Key outcomes: "social good" impact on partners, many benefits for students, and some benefits on research and curriculum
- Collaboration opportunities for other "social good" programs
  - "Data Science for Social Good" can be synergistic with "AI for Social Good"
- Visit: **dsi.ubc.ca/dssg** for more details

Outline for the Short Course

1. Overview of the DSSG Program
2. **Theme 1: Transportation**
3. Theme 2: Housing and Urban Development
4. Theme 3: Disease Control and Laboratory Testing
5. Key Technical Foundation: Natural Language Processing
6. Conclusions

DSSG Theme 1: Transportation

- Complex regarding regional interaction
  - Work with the City of Surrey, but transportation organized with other cities in Metro Vancouver
- Also complex in terms of the various datasets collected
- Multi-year program:
  - 2017: public transportation
  - 2018: private vehicle green house gas (GHG) emissions
  - 2019: electric vehicle adoption

Analysis of Riders' Tweets (2017)

(by Allahdadian, Chu, Park, Qi)

- Buses is a major form of transportation in the City of Surrey
- Online posts from bus riders allow for direct and instant feedback about their transportation experiences
- Objectives:
  - How riders are distributed in the region
  - How riders move and commute through the region over a daily cycle
  - How riders feel about their experiences

Analysis of Riders' Tweets (2017)

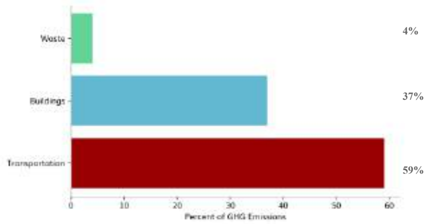
(Allahdadian, Chu, Park, Qi)

Linking data from:
- Translink trip planner and transit network
- Public twitter posts directed to @Translink while taking transit

To incorporate riders' feedback and identify new frequent transit routes

Vehicles GHG Emissions Modeling (2018)

(Anwar, Hong, Kramer and Wong)



Steps for GHG Emissions Modelling

1. Vehicle Stock Regression Modelling
   - Demographic, Spatial/Temporal elements
2. Transportation Demand Classification by vehicle class
3. Emissions Modelling

Steps for GHG Emissions Modeling: Datasets

- ICBC vehicle registration[1]
- Transportation demand model output[1]
- Building and population projections[1]
- Census / StatCan data

[1] Thank you to the City of Surrey for providing these non-open data, and students learned how to handle sensitive data

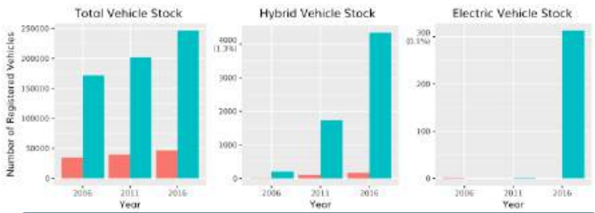
Data Integration Problem: Geographical Rebasing

ICBC Registration: **Postal Codes**

Transportation Models: **Traffic Analysis Zones**

Census: **Dissemination Area**

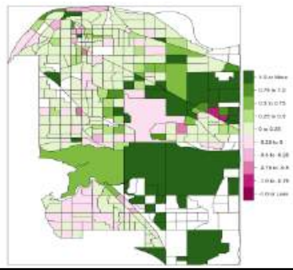
Distribution of Vehicle Stock



Subsequent analysis focused on passenger vehicle stock
(Electric vehicle analysis deferred to 2019)

Commercial Passenger

Vehicle Stock—Visualizing *Percentage Change* of Passenger Vehicles Per Capita Between 2006 and 2016



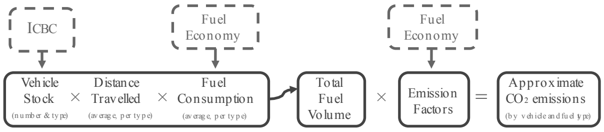
Policies regarding Neighbourhood Differences e.g., parking space, Metering

Vehicle Stock—Changes in Vehicle Attributes

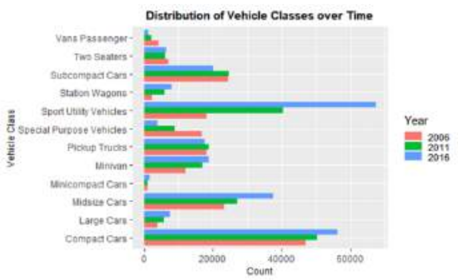
Vehicle per Capita	Vehicle Weight	Vehicle Age	TAZ Count	% of TAZ Count	% of TAZ by Pop. (in 2016)
↑ Vehicles Per Capita	↑ Weight	Older	195	52.14%	69.43%
		Younger	43	11.50%	11.92%
	↓ Weight	Older	5	1.34%	0.28%
		Younger	3	0.80%	0.99%
↓ Vehicles Per Capita	↑ Weight	Older	45	12.03%	9.25%
		Younger	11	2.94%	2.97%
	↓ Weight	Older	4	1.07%	0.23%
		Younger	2	0.53%	0.34%

Triple bad news for GHG emissions: higher vehicles per capita, heavier and older vehicles

GHG Emission Modelling Framework



Vehicle Classification—Distribution in ICBC Registry



Vehicle Stock Forecasting

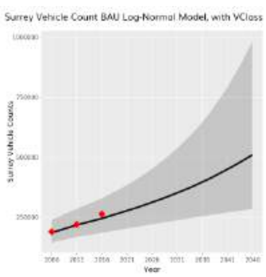
Goal

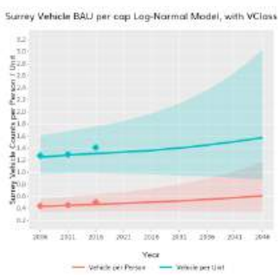
- Provide Business-As-Usual (BAU) vehicle stock size forecasts for the City of Surrey beyond the year of 2016

Challenges

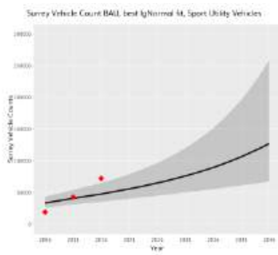
- Need to account for neighbourhood effects and population growth
- Need to account for effects of vehicle class
- Limited dataset size (only 3 time points)
  - Time series methods infeasible

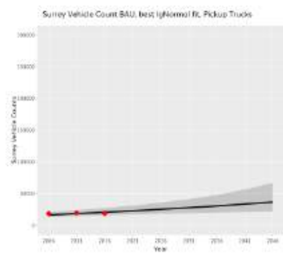
Vehicle Stock Forecasting



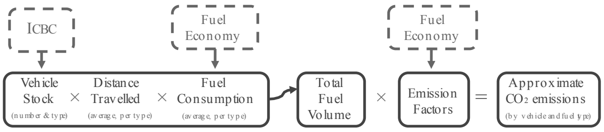


Vehicle Stock Forecasting by Vehicle Class





Producing Future GHG Emission Estimates



Surrey Electric Vehicle Project (2019)

(Greenstreet, Lai, Rodriguez-Arelis)

The Problem:
Low EV Adoption

Electric vehicle (EV) adoption challenges:

1. **Cost and models of vehicles available**
2. **Charging station Capacity**
3. Public knowledge and perception of electric vehicles

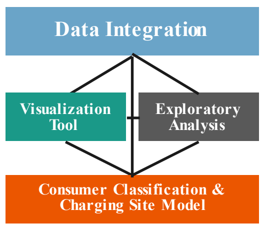
The Goal:
Help Surrey Adopt EV Faster

Provide **insights** to guide the EV strategy development

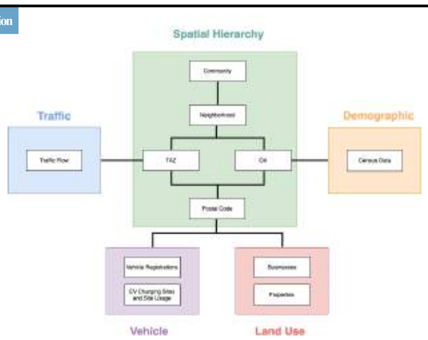
1. **Who** are the current/potential EV consumers?
2. **Where** are the potential EV consumers?
3. **How** could the city strategically put the future charging sites?

Their Approach:
Enable Data-Driven Decision Making

A **database**, a web **app**, proper statistical **modelling**, and a couple of **stories**



Data Integration

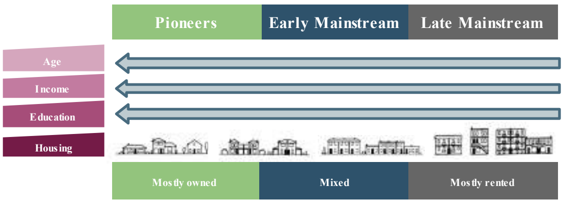


Classifying Potential EV Buyers

- Existing literature identifies key demographics for market segmentation
- Previous studies divided electric vehicle buyers into three categories:

*Canadian Zero-Emissions Vehicle Survey: Metro Vancouver Analysis (Axsen et al, 2017).

A View on Demographic Differences

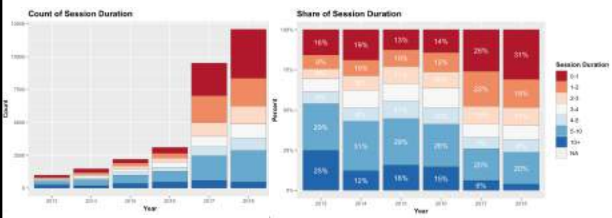


The Goal: Help Surrey Adopt EV Faster

Provide **insights** to guide the EV strategy development

1. **Who** are the current/potential EV consumers?
2. **Where** are the potential EV consumers?
3. **How** could the city strategically put the future charging sites?

Charging Session Duration in Hours



Conclusion: Transportation

- Transportation planning hugely important to a city/region
- Complexity requires multiple datasets to be used together
  - Social media data for riders' feedback
  - Census data for demographics
  - ICBC data for vehicle registration, …
- Data integration not always straightforward
- Various data science tools used
  - Sentiment analysis and natural language processing
  - Data visualization, often spatial
  - Time series forecasting
  - Classification and machine learning

Discussion: Transportation

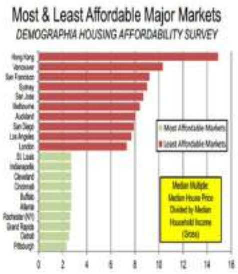
- Think about Sao Paulo, or any other city you live in
- Are there similar policy issues?
- Do similar data exist? Are they accessible?
- What are the pros and cons of this kind of "smart transportation" decision making?
- Are there bias introduced?
- Are there ethical issues?

Outline for the Short Course

1. Overview of the DSSG Program
2. Theme 1: Transportation
3. **Theme 2: Housing and Urban Development**
4. Theme 3: Disease Control and Laboratory Testing
5. Key Technical Foundation: Natural Language Processing
6. Conclusions

DSSG Theme 2: Housing and Urban Development

- Housing Cost/Affordability just as important as transportation
  - Owning or renting
- Just as complex in terms of the various datasets needed
- Multi-year program:
  - 2018: (renting) characterizing the hidden rental market
  - 2019: (while building new houses) preserving biodiversity



Preserving Biodiversity in Urban Development (2019)

(Aggarwal, Miller, Smith)

What is Biodiversity?

(Wikipedia) *'Biodiversity encompasses all living species on Earth and their relationships to each other…"*

Why do we care?

We depend on a variety of organisms for our survival!

Benefits of Biodiversity

- Oxygen production
- Carbon sequestration
- Climate regulation
- Crop pollination
- Control of soil erosion
- And more!

Key Threat to maintaining biodiversity

Habitat destruction through land use decisions, typically housing

What is the Problem?

Imagine you are an urban planner…

- You have many decisions to make about ***what*** must be built ***where***
- You also know protecting biodiversity is important for the well-being of your city

What is the Problem?

Imagine you are an urban planner…

- You would like to know what types of wildlife your region supports
  - Are their any species at risk in your area?
  - What types of ecosystems does your city have?
  - How are species distributed over time and space?

What is the Problem?

- Answering these questions by linking data from multiple sources:
  - Species occurrence data
  - Species at risk (International, National, Province)
  - What are the current gaps in the data?

Data Sources

GBIF
- About 200 data sources
  - Largest amount of observations from Citizen Science Projects
- More than 3 million species observations
- Spatial & temporal
- Taxonomy (Evolutionary tree)

Sensitive Ecosystem Inventory
- Ecosystem classification for Vancouver
- Identifies different ecosystems (Mature forest, Wetlands etc.) across the region and indicates habitat quality

Data Sources

IUCN Red List
- Database of endangered species at international level

BC Red/Blue List
- Database of endangered species at provincial level

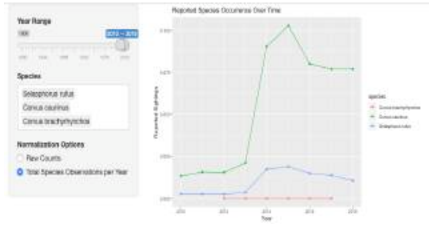
ITIS
- Provides standardized species common names and taxonomy information

Mapping species occurrence

Highlighting Endangered Species

Mapping Sensitive Ecosystems

Occurrence Trends Over Time



Big Data ≠ Sufficient Data

- Biased nature of occurrence data
  - Species (some organisms are poorly represented)
  - Time (older years have less data)
  - Space (some regions are poorly represented)
- Depicting change over time can be misleading

The Hidden Rental Market in Surrey (2018)

(Fink, Jiang, Lee, Park)

- Surrey is growing at a rapid rate
- Rental unit registration for Surrey is incomplete
- Many landlords illegally rent out secondary suites without registration
- Social consequences:
  - School overpopulation
  - Inadequate public transportation availability
  - Lack of available street parking
  - Unsafe secondary suite rentals
- Goal: provide the City of Surrey with up to date information on the **type**, **distribution** and **number** of secondary suites

Data Sources

Open Sources:

Statistics Canada

OPEN DATA

CMHC SCHL

Non-Open Sources:

SURREY

craigslist

VRBO

airbnb

kijiji

QUANTITATIVE RHETORIC

Collecting Data from Social Networking Sites

- Different web crawlers built for different websites:
  - Most postings from Craigslist: **3,000~4,000 raw data monthly**
  - Other sources (mainly Kijiji and VRBO) comprise ~300 data monthly
  - Short-term rental very few: VRBO and Airbnb
- Crawlers deployed on UBC server and collected data every day
- Current research was mainly based on data collected over 3 months

Data Cleansing & Pre-Processing

- Standardization

description	housing_type	lat	long	location	price	source	title
	/ 2br - 550ft			(Ladner)	$1,600	Craigslist	2 bedroom s
	/ 4br - 2430f	49.029901	-123.07195	(delta)	$2,600	Craigslist	Delta Tsaww

- Deduplication

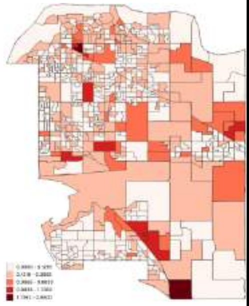
lat	long	location	price	source	title
49.188634	-122.8632	(SURREY CE	$480	Craigslist	■U+ AUG 1-FURNISHED PRIVATE ROOM - SINGLE OCCUPANCY-FEMALE STUDENT
49.188634	-122.8632	(SURREY CE	$480	Craigslist	AUG 1* BIG FURNISHED ROOM-ALL INCLUDED-SINGLE FEMALE-SURREY CENTRAL
49.188634	-122.8632	(SURREY CE	$480	Craigslist	AUG 1* CLEAN FURNISHED ROOM-ALL INCLUDED-SINGLE FEMALE-SURREY CENTRAL

- Missing value imputation for supervised-learning

lat	long	location	price	source	title
49.108673	-122.84547			Craigslist	1 bedroom in 2 bedroom basement only for girls

Spatial Distribution of Ads by Number

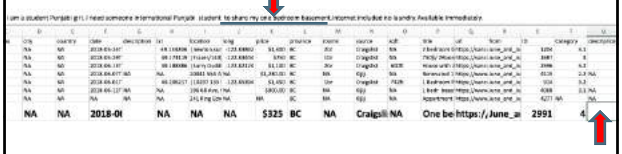
- Maps created using QGIS 3.2.3
- Counts measured using Dissemination Areas
- Highest posting densities in Douglas and City Center, high density in Cloverdale



% of online posts per neighbourhood

Unit Type Classification Example

"I am a student Punjabi girl. I need someone international Punjabi student to **share my one bedroom basement**. Internet included no laundry. Available immediately."



Building and Selecting a Classifier for Unit Type

- Manually labelled a small training set (around 200)
- Initially there were 10 unit types but finally grouped into three categories for sufficient group sizes
  - *Entire property (40%); secondary suites (40%); individual rooms (20%)*
- Used the training set to build classifiers and selected the best one
  - Naïve Bayes (75% accuracy)
  - Generalized Additive model with majority voting (83%)
  - Random Forest (91%)
- Applied the Random Forest classifier to all the 10,000+ ads

Spatial Distribution of Ads by Unit Types

Individual Room | Secondary Suite | Entire Property

Conclusion: Housing and Urban Development

- Many facets to affordable housing – from building new houses to characterizing the rental market
- Complexity requires multiple datasets to be used together
  - Census data for demographics
  - Many data sources of biodiversity including crowdsourcing sites
  - Scraping of social networking sites
- Data provenance and data cleansing important
- Various data science tools used
  - Natural language processing mainly for extraction
  - Data visualization, often spatial
  - De-duplication of tuples
  - Classification (e.g., Random forests)

Discussion: Housing and Urban Development

- Think about Sao Paulo, or any other city you live in
- Are there similar policy issues? Or are the important housing issues different?
- Do similar data exist? Are they accessible?
- Are there bias introduced?
- Are there ethical issues?

67

Outline for the Short Course

1. Overview of the DSSG Program
2. Theme 1: Transportation
3. Theme 2: Housing and Urban Development
4. **Theme 3: Disease Control and Laboratory Testing**
5. Key Technical Foundation: Natural Language Processing
6. Conclusions

68

What is the Public Health Problem?

(2018: Chen, Chiu, Lu, Zare
2019: Gao, Lam, Lee)

Lab Result
Specimen rejected | Test not performed. | No evidence of HCV infection.
No Bordetella pertussis DNA detected by PCR.
Result inconclusive. | Culture results to follow. | VaricellaZo... rus| "Isolated."
"Organism identified as" |Haemophilus influenz... erotypable(non encapsulated)

Semi-structured free form text data from lab reports containing raw test results

Project Goal: Automate the classification process!

Manual classification process (expensive, slow)

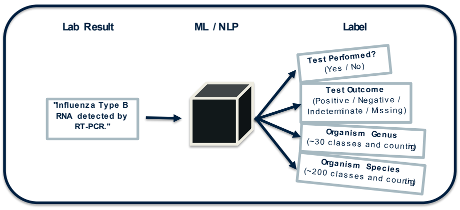
Test ...	... Outcome	Organism Name
No	Negative	"Not Found
Yes	Negative	"Not Found
Yes	Indeterminate	"Not Found
Yes	Positive	Haemophilus influenzae

Structured data used to analyze population-level disease trends

69

What is the Focus?

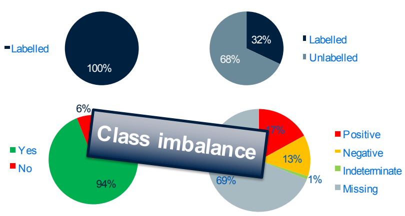
Test appropriate machine learning and natural language processing techniques for interpreting and labeling unstructured lab results, e.g., Zika, West Nile, Yellow fever, etc.



70

How does the Data Look Like?

~1 million rows; ~360K usable rows after filtering out proficiency tests and purely numeric results

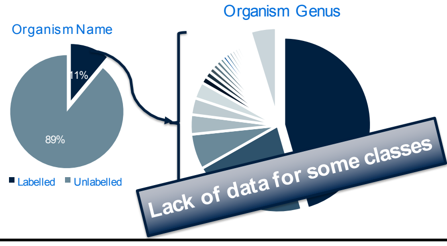


Class imbalance

71

How does the Data Look Like?

~1 million rows; ~360K usable rows after filtering out proficiency tests and purely numeric results



Lack of data for some classes

72

Most Frequent Words

Challenges Posed by the Data

- May be incomplete sentences and may contain typographical errors

BCCDC seretype: non froup 5 | Final | 12/Jun/2009 | Sputum | Streptococcus pneumoniae | STUDY
Isolate not | Salmonella species

- May contain contradictory information

TEST NOT PERFORMED | Galactomannan testing is valid only for Haematology and lung transplant patients with no recent antifungal exposure | Test performed at Provincial Laboratory of Public Health, Edmonton
Organism identified as: | Neisseria meningitidis nongroupable | Upon further investigation | Organism identified as: | **Moraxella osloensis** | by 16S rRNA gene sequence analysis.

Challenges Posed by the Data

- One organism may be positive, while another may be negative

NEGATIVE for Shiga toxin stx1 and stx2 genes by PCR. | Isolate serotyped as: | Escherichia coli | not | O157:H7

- Lots of negative organisms may be mentioned

Rhinovirus or **Enterovirus** detected by multiplex NAT. | | **Adenovirus** detected by multiplex NAT. | | Multiplex NAT is capable of detecting Influenza A and B, Respiratory Syncytial Virus, Parainfluenza 1, 2, 3, and 4, Rhinovirus, Enterovirus, Adenovirus, Coronaviruses HKU1, NL63, OC43, and 229E, hMetapneumovirus, Bocavirus, C. pneumoniae, L. pneumophila, and M. pneumoniae. | | MULTIPLE INFECTION DETECTED

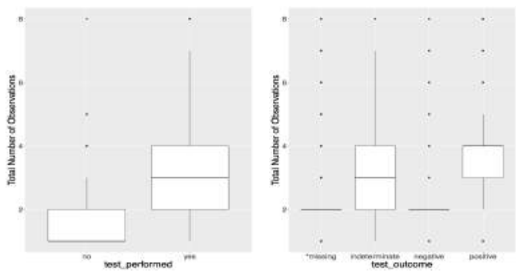
Goal: Solving Three Classification Problems

1. **Test Performed:** Yes or No
2. **Test Outcome:** Positive, Negative, Missing, or Indeterminate
3. **Organism Name:**
   - 27 classes in the labeled data
   - The classes of organisms would increase over time

Feature Engineering: Pre-processing

- Remove stop-words (e.g., 'in', 'by', 'of'), i.e., words that appear too many times
- Remove words that do not appear a minimum number of times (e.g., loquid, chacoal, laboartoy, laboraory, laboratroy)
- Then adopt a "bag of words" framework, i.e., each lab result is a bag of remaining words

Feature Engineering: Number of Observations a New Feature



Evaluating Classifier Performance: Accuracy & F-score

	Predicted Negative	Predicted Positive
Actual Negative	True Negative	False Positive
Actual Positive	False Negative	True Positive

Accuracy: (True Positive + True Negative) / Total

Precision: True Positive / Predicted Positive

Recall: True Positive / Actual Positive

F-score: (Harmonic) average of Precision and Recall

(For more than two classes, use the average F-scores for all the classes)

79

Performance of the "Test Performed" Classifier

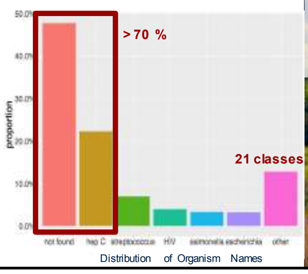
Test Performed	Bag of Words	Number of Observations
Accuracy	0.996	0.914
F-score	0.990	0.946

"Test Performed" is an easy classification problem; even a single feature can perform well

80

Dealing with Class Imbalance: Inversely weighted F-scores

- Treating all the classes with equal weights can be misleading for significantly class imbalance situations
- A model only need to perform well for large classes; poor performances for small classes are "shielded"



81

Dealing with Class Imbalance: Inversely weighted F-scores

Test Outcome	Proportion of total data	Class weight
Positive	0.17	0.06
Negative	0.13	0.08
Missing	0.69	0.02
Indeterminate	0.01	0.84

82

Performance of the "Test Outcome" Classifier

Test Outcome	Bag of Words	Number of Observations
Accuracy	0.996	0.720
F-score	0.978	0.295
Inversely-weighted F-score	0.932	0.03

- Best result selected from Naïve Bayes, Logistic Regression, Support Vector Machine and Random Forests
- "Test Outcome" still an easy enough classification problem

83

Organism Name: Meta Map for Annotation

'Respiratory syncytial virus detected by RT PCR using an assay capable of

Virus | Finding | Molecular Biology Research Technique | Lab Procedure

detecting influenza A B and RSV. Indeterminate for influenza B virus by RT PCR'

84

Using Meta Map to Extract Organism Names

Result Description	Meta Map Organism Candidates
bordetella parapertussis $ positive	"Bordetella parapertussis Bacterium"
findings equivocal for hcv infection. a follow up specimen of edta blood is requested to test for hcv rna by qualitative rt-pcr to define status of hcv infection. | hepatitis c tests completed on previous specimen.	"hcv", "hepatitis c"
no growth of salmonella	"salmonella"

85

Performance of the "Organism Name" Classifier

organism_name	Bag of Words	Bag of Words + Number of Observations	Bag of Words + MetaMap	Bag of Words + MetaMap + Number of Observations
Accuracy	0.946	0.954	0.965	**0.966**
F-score	0.751	0.788	0.865	**0.873**
Inversely weighted F-score	0.656	0.715	0.841	**0.860**

86

More Work to be done on "Organism Name"

- Finding all the organism names in a test result and their corresponding test outcomes
  - Organisms not appearing in the labeled dataset
- Will try a rule-based Meta Map approach:
  - Use Meta Map and rules to find organism names
  - Then apply the current pipeline to enhance the classifier
- Also need to deal with negation (e.g., "no growth of Salmonella")

87

Conclusions: Disease Control and Laboratory Testing

- Centre for Disease Control need to process a large number of laboratory test results
  - Particularly onerous during peak flu season, and even more problematic for disease outbreak
- Explore how to use NLP techniques to extract features and build classifiers for automated processing
  1. Test performed? (2 classes)
  2. Test outcome? (4 classes)
  3. Organism names? (many, and continuing to grow)
- First two classification problems show very good results
  - Still ethical issues about false negatives or false positives
- The last problem of organism names are harder, particularly in dealing with new organisms

88

Conclusions: DSSG Impact

- Key outcomes: "social good" impact on partners
  - Transportation
  - Housing and urban development
  - Disease control and laboratory testing
  - And more: transparency government, …
  - Visit ***dsi.ubc.ca*** for more details
- Very popular for students, project partners and sponsors
- Collaboration opportunities for other "social good" programs
  - "Data Science for Social Good" can be synergistic with "AI for Social Good"

89

Discussion: Disease Control

- Think about Sao Paulo, or any other city you live in
- Are there similar issues? Or are the important housing issues different?
- Do similar data exist? Are they accessible?
- Are there ethical issues?

- **Any comments on the DSSG program in general?**

90

Outline for the Short Course

1. Overview of the DSSG Program
2. Theme 1: Transportation
3. Theme 2: Housing and Urban Development
4. Theme 3: Disease Control and Laboratory Testing
5. **Key Technical Foundation: Natural Language Processing**
6. Conclusions

91

Common Tools in DSSG

- Data visualization, particularly map overlays
- Database tools: data cleansing, database querying
- Classification tools: e.g., random forests, support vector machine, etc.

- **Natural language processing tools:**
  - Sentiment analysis (transportation)
  - Term and topic extraction (housing, disease control)
  - Rhetorical analysis
  - Summarization

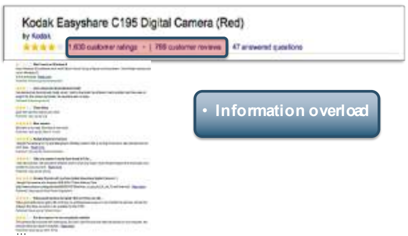
92

Motivating Application: Sentiment Analysis

Informal Documents and Evaluative Text

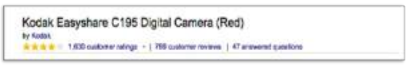
- Formal Documents: newspapers, reports, etc.
  - focus of Natural Language Processing (NLP) until the past decade
- Informal Documents: emails, blogs, MSN, user reviews, etc.
- Rapid accumulation of evaluative text
  - Expressing the author's subjective sentiments, e.g., like or not like, good or bad
- Key differences from formal text: can be very short, looser grammar, misspelling, part of a larger conversation, can be very numerous

Why Reviews Summarization



- Information overload

Extractive Reviews Summarization



"The camera is easy to use and takes clear and quality pictures."
"In addition, support for the item was no longer available, I returned the camera."

Extractive Reviews Summarization

Kodak Easyshare C195 Digital Camera (Red)

"The camera is easy to use and takes clear and quality pictures."
"In addition, support for the item was no longer available; I returned the camera."

- Useful but:
  - Lack of Coherency
  - Doesn't aggregate opinions
  - Can't express distribution of opinions

Abstractive Reviews Summarization

Kodak Easyshare C195 Digital Camera (Red)

All reviewers (45 people) who commented on the camera, thought that it was really good mainly because of the photo quality. Accordingly, about 24% of the reviewers commented about the control and they mentioned that it was fine. Also, related to the control, 7 users expressed their opinion about the auto mode and they liked it.

Abstractive Reviews Summarization

Kodak Easyshare C195 Digital Camera (Red)

All reviewers (45 people) who commented on the camera, thought that it was really good mainly because of the photo quality. Accordingly, about 24% of the reviewers commented about the control and they mentioned that it was fine. Also, related to the control, 7 users expressed their opinion about the auto mode and they liked it.

- More appropriate [Carenini 2013]:
  - More Coherent
  - Aggregates opinions
  - Can express distribution of opinions

A Sample Review

- …… the canon computer software used to download , sort , . . . is very easy to use. the only two minor issues i have with the camera are the lens cap (it is not very snug and can come off too easily). . . . the menus are easy to navigate and the buttons are easy to use. it is a fantastic camera . . . .

Extracting Evaluative Features

1. Which **features** of the entity are evaluated in the reviews?
2. What is the **polarity** of each feature? (positive or negative)
3. What is the **strength** of each feature? (rather good vs. extremely good, [-3 .. +3]

[Hu, Liu 2004; Wilson et al. 2004, Etzioni 2005]

Example: Extracting Features

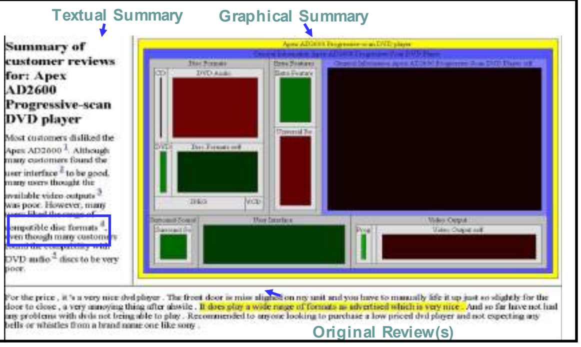
- …… the canon computer software used to download , sort , . . . is very easy to use (+2). the only two minor issues i have with the camera are the lens cap (it is not very snug (-2) and can come off too easily (-2)) . . . . the menus are easy to navigate(+1) and the buttons are easy to use(+1). it is a fantastic(+3) camera …

Grouping Extracted Features

- Map extracted features onto a taxonomy of product features at different levels of abstraction
- Such a mapping:
  - Eliminates redundancy
  - Provides a conceptual organization of the features
  - Increases user familiarity with the features

Aggregating Extracted Sentiments

- Digital Camera **[-1,-1,+1,+2,+2,+3,+3,+3]**
- 1. User Interface **[+2]**
  - Button **[+1]**
  - Menus **[+2,+2,+2,+3+3]**
  - Lever **[ ]**
- 2. Convenience **[ ]**
  - Battery **[ ]**
    - Battery life **[-1,-1,-2]**
    - Battery charging system **[ ]**
  - ….
- 3. ….



Emerging Sentiment Application: Monitoring Patients from Homes

- Beyond the typical physiologic data collected by sensors, ***text stream*** data were also collected for stay-home patients
  - Patients communicating with family caregivers, and
  - Patients chatting with fellow patients in a secured social media forum
- Exploring whether text analytics can be applied to mine such data

Text Analytics for Early Onset of Dementia [Carenini16]

- Patients were asked to describe a given picture
- Answers transcribed including stutters, false starts, and filled pauses ("um", "ah")
- Extracted 136 lexical features (e.g., vocabulary richness, information content, repetitiveness)
- 526 dementia patients vs 557 control patients
- Sensitivity = 87%; specificity = 85%

Text Analytics for Chronic Disease Management

- Text data are there, e.g., whatsapp, clinical trials
- Patients describing their own sentiments – a "window" into their psychological states, their cognitive states, etc.
- *Longitudinal* text – capturing changes over time -can be the basis of a powerful predictive model
- Building a predictive model using text to monitor patients is an important emerging area

Topic Modeling (Social Media)

Simplest Topic Modeling

- Given a collection of documents, we want to identify a list of topics covered by the documents
- Frequency-based: Word Cloud
- Deeper modeling
  - Segmentation: assigning the sentences to topics
  - Labeling: creating a natural language description of the topics

Email Example: Segmentation

- *From:* Charles *To:* WAI AU Guidelines *Date:* Thu May *Subj:* Phone connection to ftof meeting.
- It is probable that we can arrange a telephone connection, to call in via a US bridge.
- <Topic id =1>
- Are there people who are unable to make the face to face meeting, but would like us to have this facility?
- <Topic id =1>
- ……….
- *From:* Charles *To:* WAI AU Guidelines *Date:* Mon Jun *Subj:* RE: Phone connection to ftof meeting.
- Please note the time zone difference, and if you intend to only be there for part of the time let us know which part of the time.
- <Topic id =2>
- 9am - 5pm Amsterdam time is 3am - 11am US Eastern time which is midnight to 8am pacific time.
- <Topic id =2>
- Until now we have got 12 people who want to have a ptop connection.
- <Topic id =1>

Email Example: Labeling

- *From:* Charles *To:* WAI AU Guidelines *Date:* Thu May *Subj:* Phone connection to ftof meeting.
- It is probable that we can arrange a telephone connection, to call in via a US bridge.
- <telephone connection>
- Are there people who are unable to make the face to face meeting, but would like us to have this facility?
- <telephone connection>
- ……….
- *From:* Charles *To:* WAI AU Guidelines *Date:* Mon Jun *Subj:* RE: Phone connection to ftof meeting.
- Please note the time zone difference, and if you intend to only be there for part of the time let us know which part of the time.
- <time-zone difference>
- 9am - 5pm Amsterdam time is 3am - 11am US Eastern time which is midnight to 8am pacific time.
- <time-zone difference>
- Until now we have got 12 people who want to have a ptop connection.
- <telephone connection>

Topic Modeling for Documents

- Applications:
  - Information Extraction
  - Conversation visualization
  - Summarization
- Let us first consider the non-Markov version, which we call the bag-of-words topic modeling, or *Latent Dirichlet Allocation (LDA)*
- A model that "generates" D documents in a corpus covering K topics (K an input parameter given)

LDA

- Each topic *i* is described by a multinomial distribution of words, e.g., β_i = (research 0.3, support 0.3 , grant 0.2, acknowledgements 0.2, vector 0, machine 0)
- A document *d* is a bag of words with a multinomial distribution over the K topics, e.g., θ_d = (topic1 0, topic 2 0.5, topic3 0.5, topic4 0)
- A document is generated/modeled as:
  - Pick a topic distribution θ_d
  - for each word in the document, pick a topic *i* based on θ_d, and then use β_i to draw the word for the document

LDA Parameter Learning by EM

- Given a collection of documents, all the parameters of the LDA are solved by using EM [Blei 2003]
- *Expectation-Maximization* strategy well known for learning latent variables and model parameters
  - Iterate between topic descriptions and each document's distribution of topics
  - Log likelihoods improve from one iteration to the next until "convergence"

E.g., Newspaper articles

One Improvement of LDA

- A document is an ordered sequence of words; yet LDA completely ignores the ordering of words
- To improve modeling, we can impose a Markov chain on the latent variables
- ψ_j equals 1 if there is a topic change, 0 otherwise
  - If ψ_j = 0 across all the words in a document, we restrict one topic per document
  - If ψ_j = 0 across all the words in a sentence, we restrict one topic per sentence
  - The LDA model essentially allows a potential topic change per word

Extensions of LDA

- Not Requiring the number of topics, K, to be known a prioi
  - Dealing with a new document on an unseen topic
- Incorporating known correlation (or lack thereof) between words, e.g., heart failure, blood pressure
- Dealing with meta-data, e.g., author, date
- Even optimized for other types of data, e.g., genomics, images, conversations [Blei 2012]

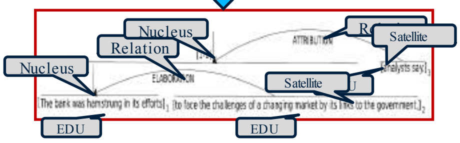
Rhetorical Analysis and Parsing

What is Rhetorical Structure?

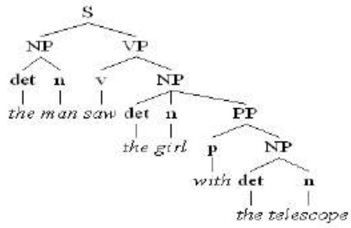
- ***Rhetorical relation***: *"is a description of how two segments of discourse are rhetorically connected to one another"*
- ***Rhetorical structure*** is a description of the rhetorical relationships among different parts of a discourse
- E.g., rhetorical parse tree of a document - both intra-sentential and inter-sentential
- A lot more semantical than dependency parsing, which is more syntactical

E.g., *Intra*-sentential Rhetorical Parse Tree

The bank was hamstrung in its efforts to face the challenges of a changing market by its links to the government, analysts say.



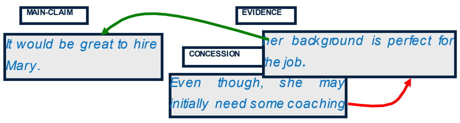
Different from Standard Dependency Parsing



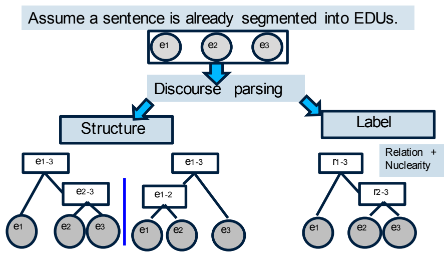
Discourse Parsing: Example

What is the main claim in a message and how it is expanded/supported by the other claims

"It would be great to hire Mary. Even though she may initially need some coaching , her background is perfect for the job."



Intra-sentential Discourse Parsing



A CRF-based Discourse Parser [Carenini 2012]

Discourse Parsing State of the art limitations:

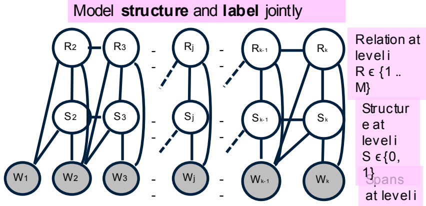
- Structure and labels determined separately
- Do not consider sequential dependency
- Suboptimal algorithm to build structure

Their parser addresses these limitations

- Layered ***Conditional Random Fields***
- A generalization of HMM, loosening directionality of the edges in the graph

The Parsing Model

Model **structure** and **label** jointly

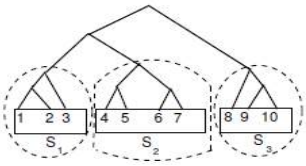


Dynamic Conditional Random Field (DCRF) [Sutton-2007]

Models sequential dependencies

Multi-sentential Parsing

- Why not used the same model as for intra-sentential?
- **Not scalable**: $O(n^3)$ where n the number of sentences
- Observe: most sentences have well-defined parse trees
- Why not build trees for sentences first, then build on top of those?



Rhetorical Parsing Enhancing Abstractive Sentiment Summarization

Recall: Why Reviews Summarization

- Information overload

Recall: Extractive Reviews Summarization

"The camera is easy to use and takes clear and quality pictures."
"In addition, support for the item was no longer available, I returned the camera."

Recall: Extractive Reviews Summarization

"The camera is easy to use and takes clear and quality pictures."
"In addition, support for the item was no longer available, I returned the camera."

- Useful but:
  - Lack of Coherency
  - Doesn't aggregate opinions
  - Can't express distribution of opinions

Recall: Abstractive Reviews Summarization

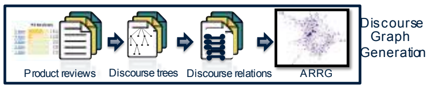
All reviewers (45 people) who commented on the camera, thought that it was really good mainly because of the photo quality. Accordingly, about 24% of the reviewers commented about the control and they mentioned that it was fine. Also, related to the control, 7 users expressed their opinion about the auto mode and they liked it.

Recall: Abstractive Reviews Summarization

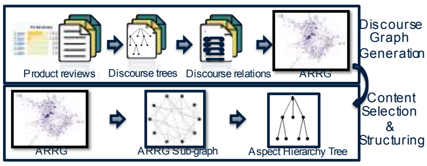
All reviewers (45 people) who commented on the camera, thought that it was really good mainly because of the photo quality. Accordingly, about 24% of the reviewers commented about the control and they mentioned that it was fine. Also, related to the control, 7 users expressed their opinion about the auto mode and they liked it.

- **More appropriate** [Carenini 2013]:
  - More Coherent
  - Aggregates opinions
  - Can express distribution of opinions

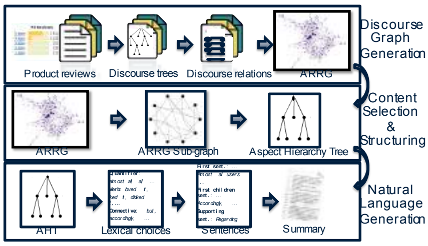
Mehdad Framework



Mehdad Framework



Mehdad Framework



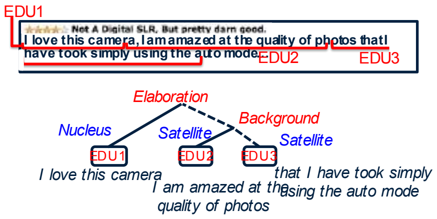
Example: Rhetorical Parsing of a Review

Not A Digital SLR, But pretty darn good.
I love this camera, I am amazed at the quality of photos that I have took simply using the auto mode...

Example: Rhetorical Parsing of a Review

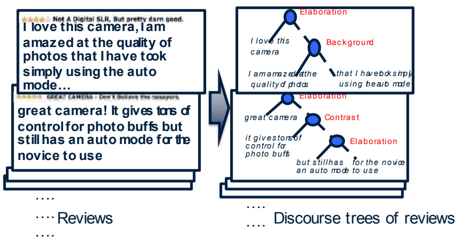
EDU1
Not A Digital SLR, But pretty darn good.
I love this camera, I am amazed at the quality of photos that I have took simply using the auto mode. EDU2 EDU3

Example: Rhetorical Parsing of a Review

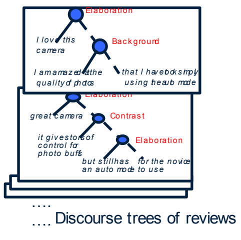


Rhetorical Parsing of All Reviews

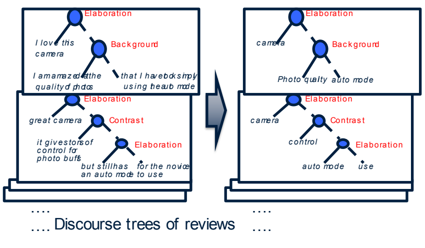
Rhetorical Parsing of All Reviews



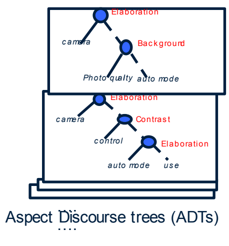
Aspect-based Discourse Tree (ADT)



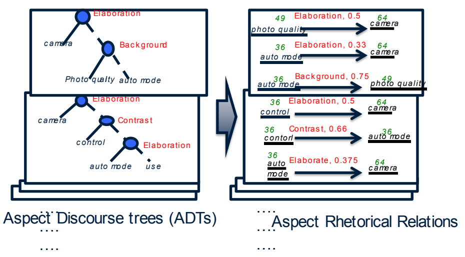
Aspect-based Discourse Tree (ADT)



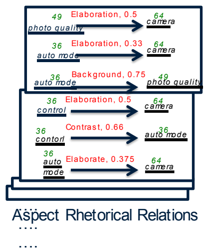
Aspect Rhetorical Relations



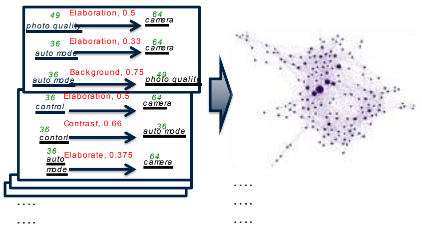
Aspect Rhetorical Relations



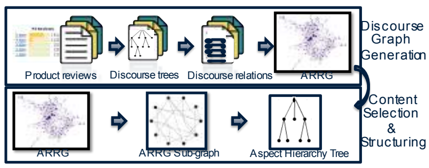
Aspect Rhetorical Relation Graph



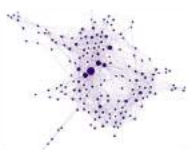
Aspect Rhetorical Relation Graph



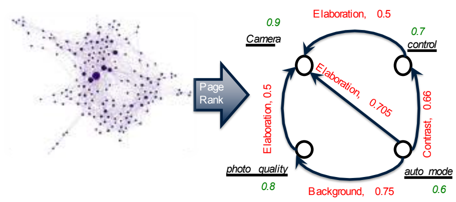
Mehdad Framework



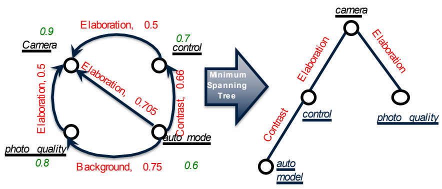
Content Selection: Subgraph Extraction



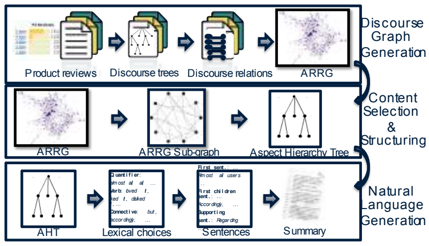
Content Selection: Subgraph Extraction



Content Structuring
Aspect Hierarchical Tree



Mehdad Framework



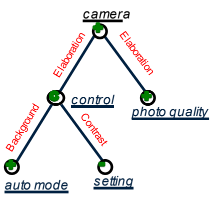
Microplanning (Lexical Choice)

Lexical Choice	Case	Examples
Quantifiers	Relative number of people	*"Almost all users", "About half of the users", "n users", "Around y% of the shoppers"*
Polarity Verb	Mixed opinions, Average polarity	*"expressed controversial opinions about this feature", "loved it", "disliked it"*
Connectives	Polarity agreements, Relation, Number of children	*"Also, related to the aspect", "Accordingly, ", "similarly", "In contrast"*

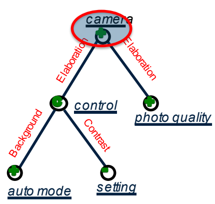
Sentence Realization

Sentence Templates	Case	Template Example
First Sentence Template	Polarity agreement, Relation	*"quantifier + polarity-verb + 'mainly because of the' + highest-weighted-child", "quantifier + polarity-verb"*
First level children templates		*"connective + ',' + quantifier + polarity-verb"*
Supporting Sentences Template	Number of children, Polarity agreement	*"connective + quantifier + verb", "connective + quantifier + verb + [and, similarly, while] + quantifier + verb" "connective + quantifier + verb + [but, in contrast, on contrary] + quantifier + verb"*

Example: Sentence Realization

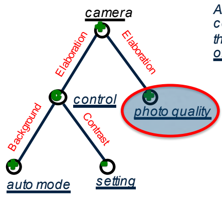


Example: Sentence Realization



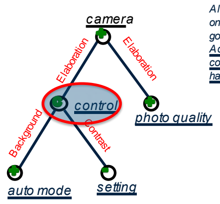
All reviewers (45 people) who commented on the camera thought that it was really good

Example: Sentence Realization



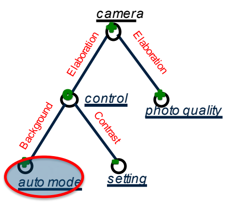
All reviewers (45 people) who commented on the camera, thought that it was really good mainly because of the photo quality.

Example: Sentence Realization



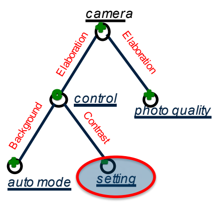
All reviewers (45 people) who commented on the camera, thought that it was really good mainly because of the photo quality. Accordingly, about 24% of the reviewers commented about the control and they had mixed opinions about it.

Example: Sentence Realization



All reviewers (45 people) who commented on the camera, thought that it was really good mainly because of the photo quality. Accordingly, about 24% of the reviewers commented about the control and they had mixed opinions about it. Also, related to the control, 7 users expressed their opinion about the auto mode and they liked it,

Example: Sentence Realization



All reviewers (45 people) who commented on the camera, thought that it was really good mainly because of the photo quality. Accordingly, about 24% of the reviewers commented about the control and they had mixed opinions about it. Also, related to the control, 7 users expressed their opinion about the auto mode and they liked it, in contrast, 6 shoppers commented about the setting and they didn't like it.

Summary: NLP Background

- Sentiment analysis (transportation)
  - Extraction of features, polarity and strength
  - Visualization
- Term and topic extraction (housing, disease control)
  - LDA a breakthrough algorithm with lots of extensions
- Rhetorical analysis
  - Identification of rhetorical relations
  - Extractions and their uses made more accurate
- Summarization
  - Abstractive vs extractive summarization
  - Abstractive summarization of sentiments

Conclusions: DSSG

- Vision: To facilitate a quantum leap in society's ability to gain value from data by
  - enhancing the capability of students to learn from data
  - inspiring many more students towards further study of data science, and
  - engaging partners to work on projects with high societal value
- "Social good" impact on partners
  - Transportation
  - Housing and urban development
  - Disease control and laboratory testing
  - And more: transparency government,…

Thank you for inviting me!

rng@cs.ubc.ca
dsi.ubc.ca

Thank You