Semantic Text Classification Research with Industrial Objectives– The Story of Scamseek

Prof Jon Patrick

Sydney Language Technology Research Group

School of Information Technologies

University of Sydney

The University of Sydney

Presentation Structure

- The Nature of Text
- The Task Definition
- The Requirements for TC
- Semantic vs. Lexical solutions
- Scamseek Operationally
- Scamseek Software
- ScamSeek Results

Scheme of Motivation

- Surveillance Diagram

The Nature of Text

- Conceptual Reach
- Structural Reach
- Semantic Reach

- Meaning as a System

Conceptual Reach is the description of the range of topic information at a coarse level which amounts to the skin of a classification.

It holds together the bones and organs of the class body and is the most visible, masquerading as the whole being while hiding behind its façade the true depth and complexity of the being.

It is typically represented by the unigrams and like all beauty is only skin deep.

Conceptual reach is the only aspect computed by current unigram approaches to text classification.

Structural Reach is the skeleton of the classification that forms the backbone of a more elegant creature.

It is ubiquitous and imperceptibly variant from body to body but remains forever necessary without ever being consciously recognised.

Yet its variability contributes secretly to the shape of each class family in significantly imperceptible variation from one to another.

It is represented by the syntax of the delivery language.

Semantic Reach is the true depth of the class. It is the visceral part that represents the entity in its ultimate complexity.

It is made up of many soft and flexible parts, forever omnipresent and controlling the functioning but yet only visible through the very special lenses of close and detailed linguistic modelling.

It is the most complicated part of the body hidden yet contained by the skin but forever clinging to the form created by the structural skeleton.

Its interconnectivity is infinitely richer than its container and its architectural frame but even more hidden than either.

Its richness is what makes it the hardest part of the system to perceive, catalogue and compute.

It is only identified by deep semantic analysis of the lexico-grammar of the texts.

Systemic Cohesion is the integration of Conceptual, Structural and Semantic Reach. It captures the essence of the inner dynamism of the body parts and also how they interact with each other and how the whole body interacts with the world around it.

Systemic Cohesion understands that the complete meaning of the text is represented not by only the components used for each type of reach but by the choices available in each case and the mental decisions made by the speakers or writers to create their representation.

Meaning making is seen as organic and a matter of choices made from the language repertoire and these choices are meant to serve in conveying the intended meanings in the text.

Linguistics

- Create corpus
- Identify text types
- Identify class memberships
- Characterise class profiles
- Sub-divide classes into registers
- Characterise registers
- Quality control – feedback cycles

Task Definition

- What is the corpus?
- Where is the corpus?
- What is the classification based on?
- Types of classifiers?
- System for Forensics or Surveillance?
- What is a text?

1.1 Constructing a Corpus

- Harvest the texts
- Store the texts
- Review/Clean the texts
- Construct a categorisation scheme
- Read texts
- Assign texts to their class

1.2 Classification of documents

- Set criteria for classes
- Set criteria for sub-classes/registers
- Validate assignments with a second classifier
- Partition an audit corpus

1.3 Methods of Corpus Building

- Supplied or need to Harvest
- Co-train or manual classification
- Constraint train- accumulate the most likely texts by selective co-training and manual classification (active learning)

Designing the Classification Scheme

- Selecting classes
- Assigning class membership
- Selecting & Assigning sub-classes
- Class membership revision
  - Impact on annotation & storage
  - Impact on current experiment results

Texts of Interest

- Documents
- Chat
- SMS
- Bulletin Boards
- Web Pages
- Essays

Constructing a language model-features?

- Bag-of-words, unigrams
- N-grams
- Chunks
- Syntactic structures
- Metadata
- Semantics
- Pragmatics

Conceptual (N-gram) model of text

- Words or selected word groups are tokens
- Tokens are counted
- Counts represent the underlying distribution of word usage in language and therefore represent the differences in the classification scheme

Structural Model of text

- Metadata
- N-grams
- Contained corpus
- Contained classification

Semantic model of text

- Semantics is
  - A matter of word selection – n-grams
  - A matter of word meaning -WSD
  - A matter of syntactic structure
  - A matter of metadata

Systemic (Functional) Model of Text

- Meaning is a matter of choice to achieve
  - A deliberative interpersonal position
  - A message of ideas to be conveyed
  - All delivered from a preference drawn from a range of structural choices

Can the models be fused?

- Fusing the models is valuable as it
  - potentially offers a means of differential granular analysis
  - Allows current computational methods to take a place at the solution table
  - Minimises work on least relevant material
  - Maximises people availability for most relevant materials

Computing With Text

- Language Modelling
- Language (Pre-)Processing
- Feature Selection
- Attribute construction
- Classifier Model Selection
- Classifier Inference
- Engineering a client solution

Language Processing Strategy

- Tokenisation
- Multi-word items
- Entities
- Parsing
- Ontologies

What is the feature representation- Attributes?

- Features are the conceptual elements you wish to assess
- Attributes are mapping of features to numeric data representation
- Mappings are many and varied and take in account many characteristics of the text processing problem

Classifier Development

- Feature selection
- Mapping to attributes
- Methods of attribute selection
- Program of Experimentation
- Optimising class-register configurations

Software Engineering

- Systems
  - Harvesting
  - Data Management
  - Experiment Code Management
  - Service Code Management
  - Production System Generation

Project Management

- Scheduling milestones
- Costing
- Recruitment
- Setting client expectations
- Reaching objectives
- Delivering to deadlines & specifications
- Setting future plans

Research & Development in the one Project

- Doing both at the one time
  - Needs clear separation of objectives
  - Needs co-ordination of each team role
  - Needs belief in the value of both
  - Only works when final processing functionality is non-determined
  - Requires planning for both types of work
  - Has fantastic advantages

The Features of the Scamseek Technology

- Separates texts with subtle differences
- Finds text classes of very low frequency in a collection
- Finds very small texts of interest
- Uses the meaning intentions of the authors – not the word strings
- Developed on principled grounds of linguistics, computational theory & software engineering
- Shown to work effectively for a wide range of financial scams

ASIC - Financial Scams

- Illegal Offerings
- Unlicensed Advisors
- Share Ramping

ASIC's Surveillance Context

- Internet Surveillance Challenge
- Surf Days
  - 30 people for a day, every quarter
  - 5,000 docs vetted
  - Reduced successively to 1500, 200, 50, 20
- 1st Solution – Webhound – string searches

ScamSeek – Phase 1

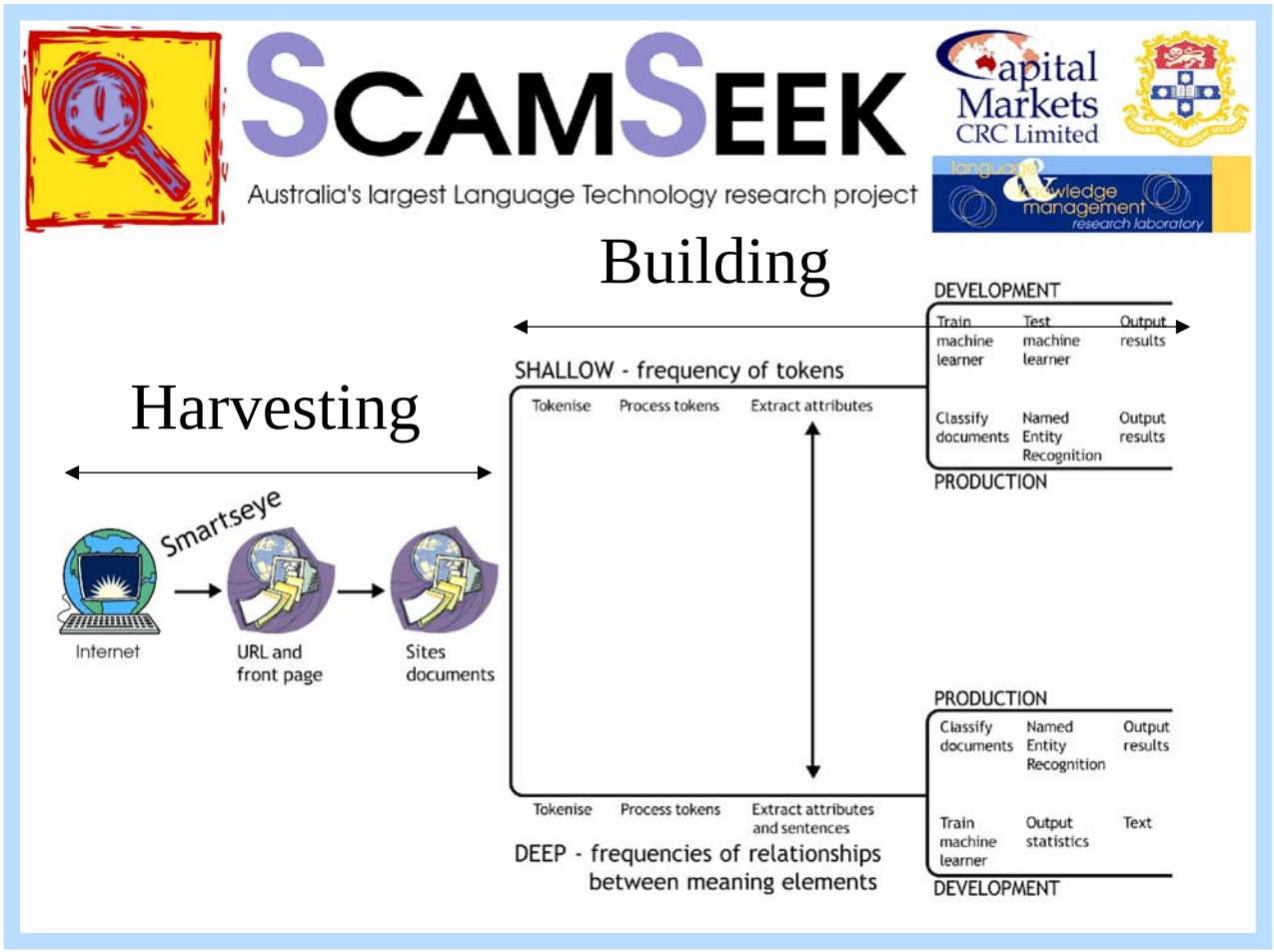
- $1m budget, 6 months elapsed time
- Document Classification Task
- Joint Funding by ASIC, CMCRC, U of Sydney, Macquarie U
- System installed September 2003
- First litigation currently in court : Grammax case
- 8 referrals made to overseas agencies

ScamSeek Phase 2
Oct 2003-June 2004

- Improve results for Web Pages
- Develop Classifiers for other Internet data types

Challenges

- Use more sophisticated linguistics
- Build harvesters where needed
- Collect corpora
- Identify scams
- Analyse linguistics

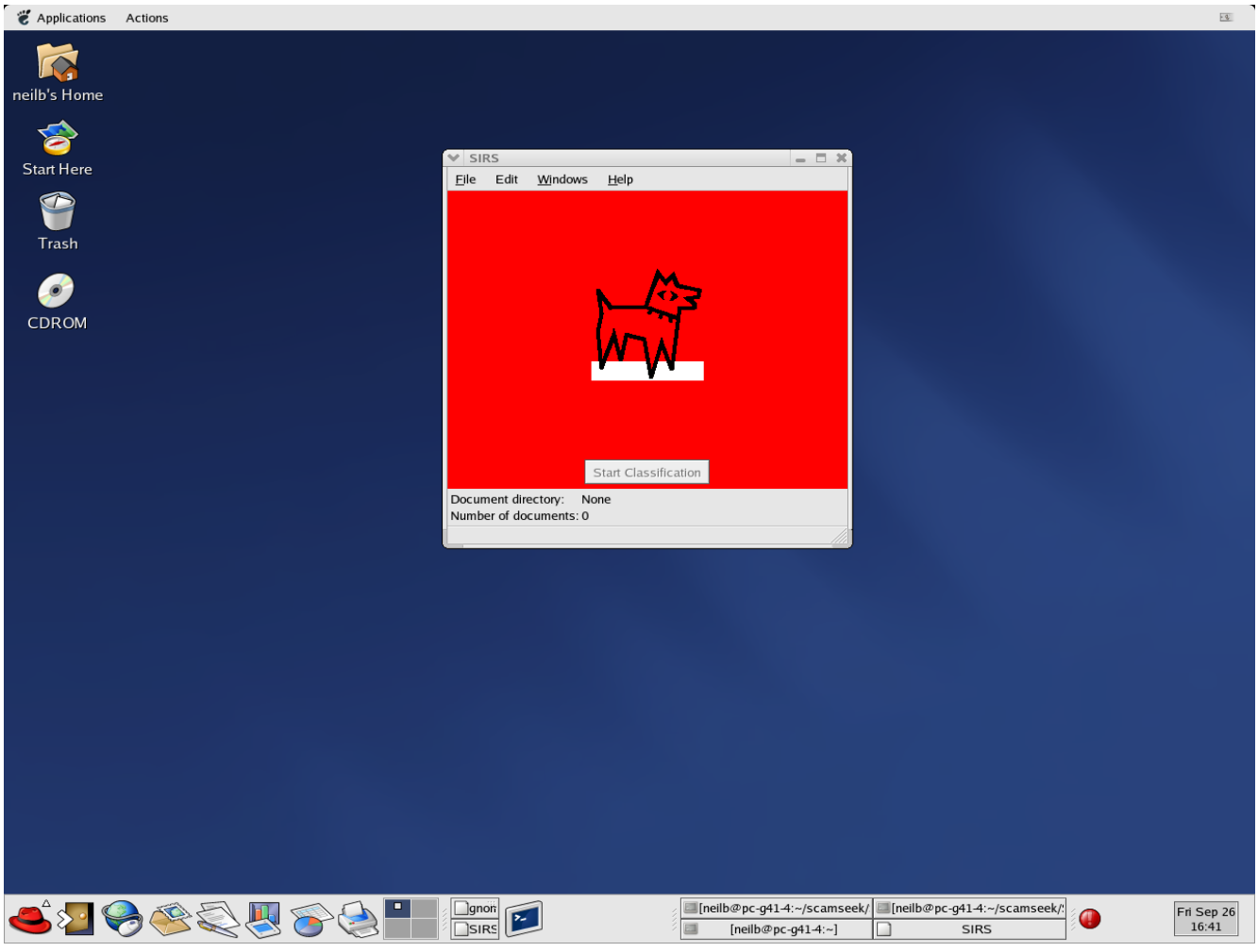


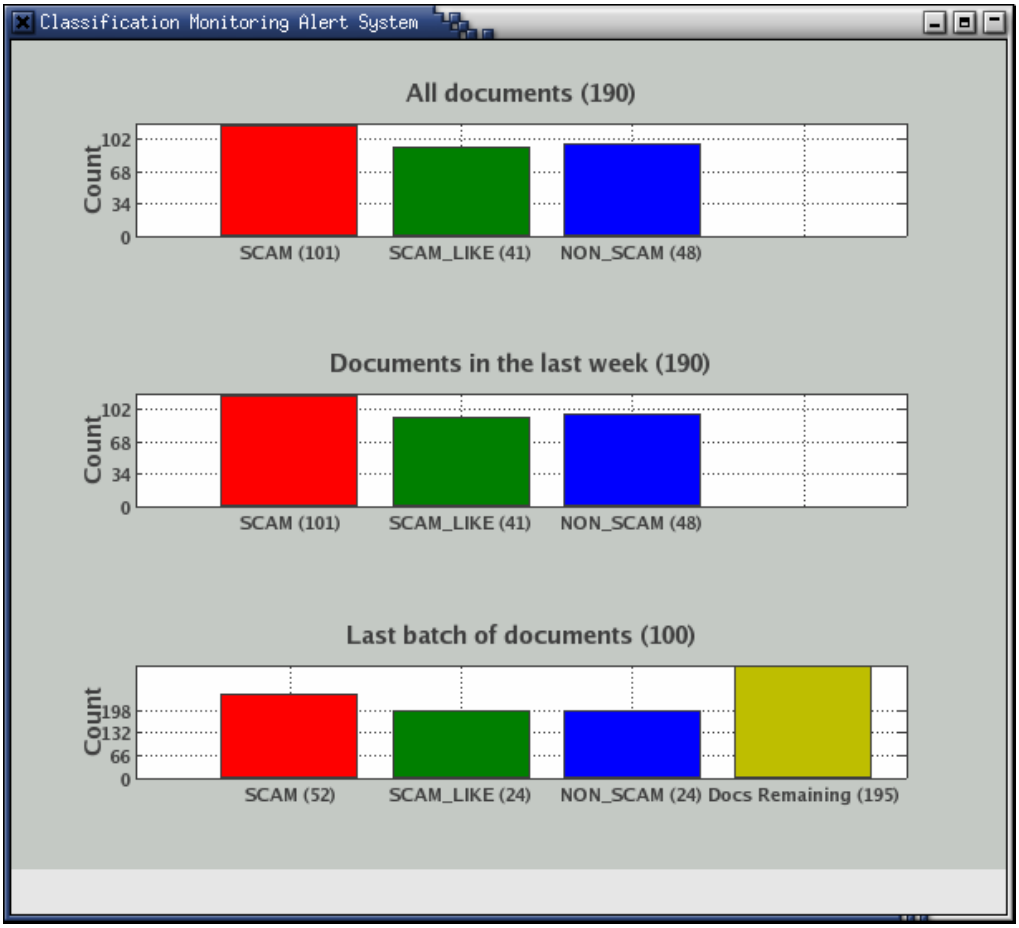
Organisational Processes

- Client team – create specifications
- Linguists team – model the language
- Computational Linguists – customise the classifier
- Software Engineering team – deliver the working system

Open Source Software Tools

- Linux
- Python
- GTK with GLADE
- Postgres
- Weka (Machine Learning)
- Bugzilla
- CVS
- Twiki



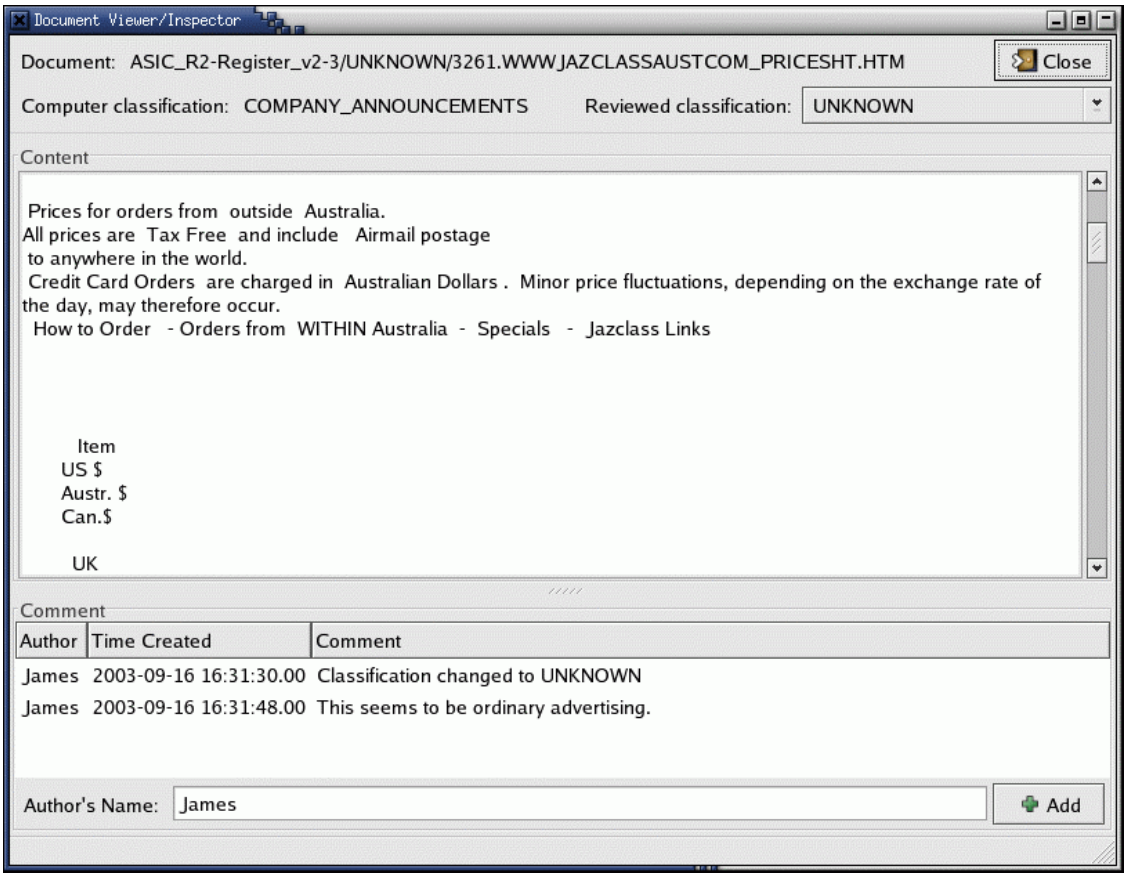


Document Classification Result

File Edit View Help

Batch	Document	Classification	Verified
Batch one (30)			
2003-09-14 14:23:26			
2003-09-15 11:12:42			
	5185.WWWZEITGEISTNETAU_PROFILE_C.HTM	OFFSHORE_ENTITY	
	1932.WWWCBCCA_REUTER_20030205_AI.HTM	COMPANY_ANNOUNCEMENTS	
	2370.WWWICAAUSTRALIACOM_ICA_NEW.HTM	MEDIA_REPORT	
	3230.WWWISTPMURDOCHEDUAU_SU_GNAR.HTM	NIGERIAN_SCAM	
	6517.WWWEXPATFOCUSCOM_VIEWFORUMP.HTM	MEDIA_REPORT	
	4318.WWWSWANMPORG_WELFARE_REFOR.HTM	UNKNOWN	
	4056.WWWROYALSOCACUK_ROYALSOC_F.HTM	NIGERIAN_SCAM	COMPANY_ANNOUNCEMENTS
	6063.BRWCOMAU_NEWSADMIN_STORIES2.HTM	NIGERIAN_SCAM	
	3082.WWWHOBBICOCOM_AIRPLANES_HCA.HTM	NIGERIAN_SCAM	MISC
	1438.WWWAPPEACOMAU_EDUSITE_HTML_.HTM	NIGERIAN_SCAM	MISC
	7222.WWWINVESTUKCOM_STUDENTS_HIS.HTM	NIGERIAN_SCAM	
	4772.WWWWETFEETCOM_EMPLOYER_INSI.HTM	UNKNOWN	
	3436.WWWMILLHOUSEIAGCOMAU_STAFF_.HTM	NIGERIAN_SCAM	
	1358.WWWANGELABOOTHZIPCOMAU_DEF.HTM	NIGERIAN_SCAM	
	721.PHYSICSOPENACUK_IAU46_NEWS.HTM	MEDIA_REPORT	
	1470.WWWARTSUSYDEDUAU_DEPARTS_M.HTM	NIGERIAN_SCAM	

Mock Data-only intended for demonstration



Mock data- only intended for demonstration

Performance Criteria

- Precision – the percentage of positive hits that are correct
- Recall the percentage of positives in the database that you found
- F-value – geometric mean of Precision and Recall

Semantic Processing 30/9/2003

Scams

- Precision = 74%,
- Recall = 35%,
- F=48%

Baseline 1000 single words, F=21%

Audit Results – 21/10/2003

ASIC Class \ Computed Class	Scam	Non-Scam	TOTAL
Scam	18	26	**44**
Non-Scam	6	1525	**1531**

Audit: Precision =.75, Recall=.41, F=.53

Train: Precision =.74, Recall=.35, F=.48

ScamAlert Current Status

- System narrows search space to find 4scams/5 docs.
- This is a reduction from 1 scam per 55 documents – 100-fold productivity gain
- System has established 19 types of scam
- System has correctly identified many incorrectly classified docs in training corpus
- System found 4 missed scams in audit corpus

Client Definitions – Scam Registers

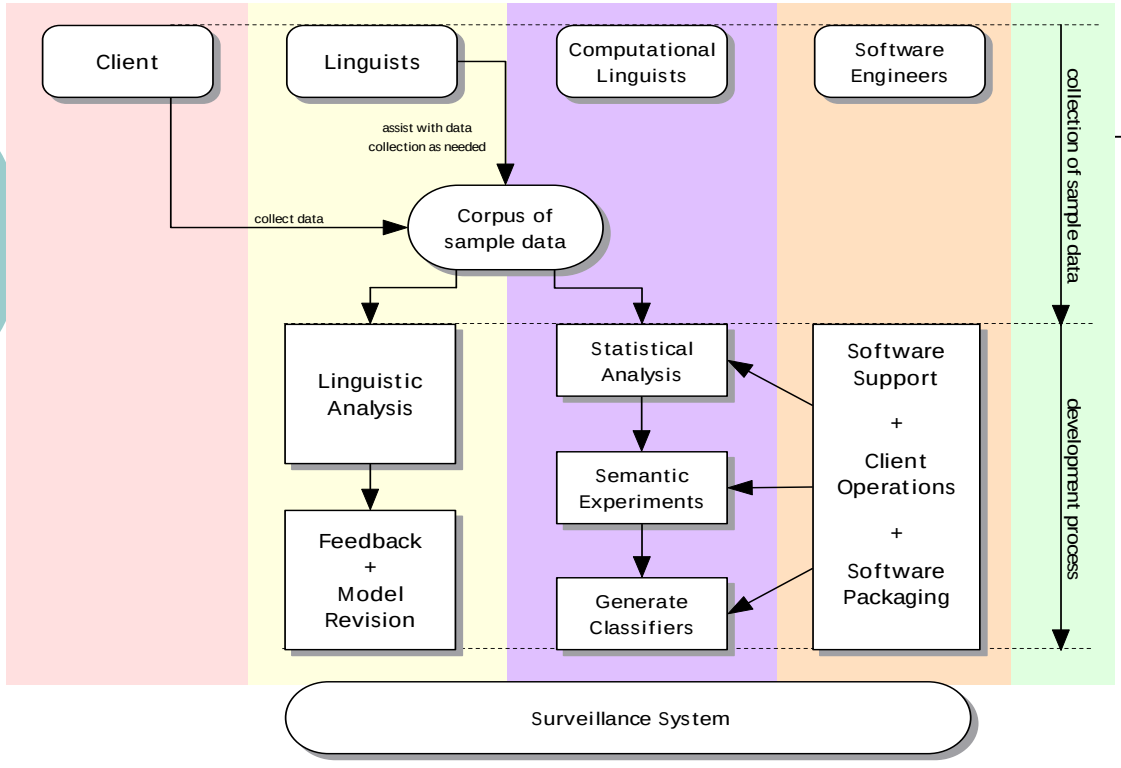
Non-ASIC Scams

Completed 30th June 2004

- Scamseek Phase 2 - $1.2m budget to run up until 30 June 2004
- Improve Web Page classifier
- Develop classifiers for other data types
- Performed beyond contract specifications
- Came in under time
- Came in under budget

Results from 3 Corpora - 30th June 2004

	Web Pages(1)	Web Pages(2)	Corpus 2	Corpus 3
Precision	.744		.850	.852
Recall	.528		.834	.639
F-value	.618		.844	.730
Scam /non-scam texts	373/6391		686/1483	1395/ 13716



Work Development Stages

- Feasibility Audit
- Trials
- Prototype
- Production System

Why Computing Meaning is Hard! or Linguists are the Lynchpin

- Same meaning is spread across classes
- Ambiguous use of words
- Multiple forms for the same meaning
- Meaning is not in the most frequent words – word does not equal meaning
- Clients don't have a precise sense of meaning

Go to Word Files

- Data Services
  - Huntley Page & Prophetnet Page
  - Usage of Terms
  - SFL Spreadsheet

Linguists Processes

- Identify strongest semantic themes in the texts
- Construct an SFL description
- Review misclassified docs
- Review optimal feature sets
- Amend SFL description/Amend document Register

Why the best Systems are Heuristic

- Impact of variables is not yet predictable
- The range of variables is large - learners, features, attributes, language extraction
- Current solutions achieved by extensive experimentation

Why Language Engineering is a Large Systems Problem

- S'ware Engineering must support 3 groups – Clients, Linguists, Comp Ling.
- Very large data volumes requiring fast access need advanced DBMSs
- System of complex interacting language processing modules
- Complex interaction between experimental and production software