Language Independent Textual Correlation Analysis

Dr Robert Sanderson
Dept. of Computer Science
University of Liverpool
azaroth@liverpool.ac.uk

Clare Llewellyn
Sydney Jones Library
University of Liverpool
clare.llewellyn@liverpool.ac.uk

http://www.cheshire3.org/
http://www.nactem.ac.uk/

Overview

- Text Mining
- National Centre for Text Mining
- Cheshire3
- Text Mining for Historians
- Association Rule Mining
- Language Independent Textual Correlation Analysis

Text Mining

What is "Text Mining":

- Commonly used definitions based on that of Data Mining:

  "The non-trivial extraction of previously unknown, interesting (facts/patterns) from a collection of texts."

- Facts: Requires semantic knowledge

  Patterns: Requires only statistical methods

  (but natural language processing also helps)

- Historical Text Mining should discover patterns or facts from historical texts, perhaps in conjunction with other sources.

Text Mining: Extraction of Facts

Extracting Facts is much harder than extracting Patterns

- Requires sophisticated natural language processing
  - Part of Speech tagging
  - Deep Parsing of phrases and clauses
  - Named Entity Recognition
  - Information Extraction
  - Information Correlation
- Sufficiently accurate tools not yet available for languages of historical texts
- Access to sufficiently large collections of appropriate historical texts not available
- Benefits would be significant in finding correlations between texts automatically

Text Mining: Extraction of Patterns

Extracting Patterns is easier but requires more human analysis

- Fundamental to Computational Linguistics
- Simple patterns of words possible with just statistics
- More complex patterns become available with natural language processing (eg: adjective* noun+)
- Further patterns with parsing of subject/verb/object in clauses (eg: “??? kills cancer” vs “cancer kills ???”)
- Tools readily available
- As discussed by Paul and Dawn!

NaCTeM: Introduction

The National Centre for Text Mining (NaCTeM) is the first publicly funded text mining centre in the world.

"It provides text mining services and tools in response to the requirements of the UK academic community. The software tools and services which NaCTeM supplies allow researchers to apply text mining techniques to problems in their areas of interest. In addition to providing services, the centre is also involved in, and makes significant contributions to, the text mining research community both nationally and internationally."

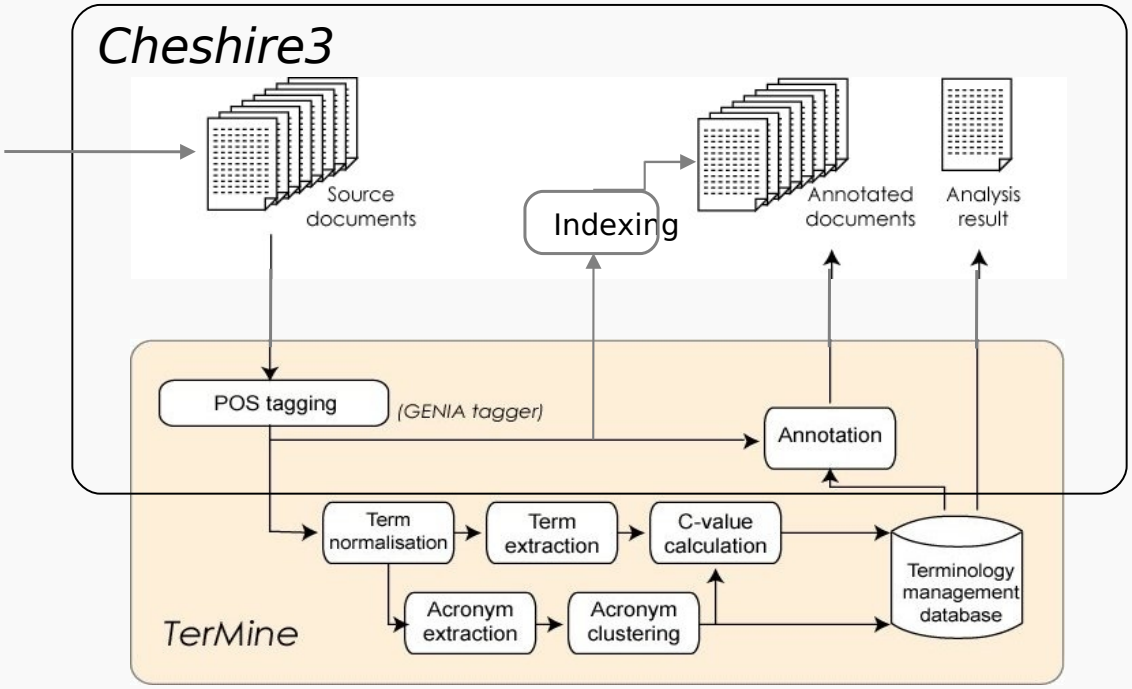
- JISC, EPSRC and BBSRC funded
- NaCTeM is operated by two universities:
  - University of Manchester
  - University of Liverpool

NaCTeM: Tools

NaCTeM is focused in the Biosciences domain.

- **TerMine** - method for automatic term recognition which finds the most important terms in a document and automatically ranks them
- **Info-PubMed** - provides information and graphical representation of biomedical interactions extracted from Medline
- **Medie** - an intelligent search engine, for semantic retrieval of sentences containing biomedical correlations from Medline abstracts.
- **AcroMine** - finds expanded forms of acronyms as used in Medline.
- **Cheshire + Termine** - a search engine for Medline documents that generates lists of significant terms relevant to the search results.

NaCTeM: Cheshire - Termine



- Improves efficiency of service through storage of pre-processed documents
- Creates ranked list of terms based on user search
- Retrieves documents based on terms
- Integrates with services and other remote providers to provide source text

NaCTeM: Cheshire - Termine Display

NaCTeM
The National Centre for Text Mining

New Search | Feedback | Contact Us

This page contains the fifty most significant terms discovered by Termine based upon all of the documents that matched your query of Medline.

- Significant Terms - A ranked list of significant terms
- Related Terms - A link to reprocess this list relative to this term
- Related Documents - The number of documents containing this term, by clicking on this the titles and links to the original article will be displayed

Significant Terms	Related Terms	Related Documents
blood pressure	terms	931
systolic blood pressure	terms	161
risk factor	terms	1859
mm Hg	terms	245
diastolic blood pressure	terms	131
heart rate	terms	444
systolic blood	terms	163
arterial blood pressure	terms	101
body mass index	terms	514
diastolic blood	terms	131
% CI	terms	1348
arterial blood	terms	153
mass index	terms	563
body mass	terms	602
cardiovascular disease	terms	482
high blood pressure	terms	51
blood flow	terms	490
mean arterial blood pressure	terms	45
heart disease	terms	501
cardiovascular risk	terms	196
hypertensive patient	terms	87
heart failure	terms	572

	PubMedId	Title	Termine
1	9886731	Determinants and significance of declining blood pressure in old age. A prospective birth cohort study.	analyze
2	9869304	Job strain and ambulatory blood pressure among female white-collar workers.	analyze
3	9864536	Incidental high blood pressure in family practice: due to hypertension and/or left ventricular hypertrophy in more than half of the patients	analyze
4	9869002	In search of hypertension genes in Dahl salt-sensitive rats.	analyze
5	9869040	Fetal carotid blood flow during videofetoscopy.	analyze
6	9894438	Stress and hypertension.	analyze
7	9869018	Antihypertensive efficacy of lercanidipine at 2.5, 5 and 10 mg in mild to moderate essential hypertensives assessed by clinic and ambulatory	analyze

Cheshire3: Introduction

History:

- Originally developed at UC Berkeley
- Solution for library data, then SGML
- Monolithic application in C + TCL

Cheshire3:

- Developed at Liverpool, plus Berkeley
- XML, Unicode, Grid scalable, Standards based
- Object Oriented, Information Analysis Framework
- Easy to develop and extend in Python
- Increasingly stable and easy to use

Cheshire3: Introduction

Environmental Requirements:

- Very Large scale information systems
  - Terabyte scale datasets (Data Grid)
  - Computationally expensive processes (Comp. Grid)
- Digital Preservation
- Analysis of data, not just retrieval (Data/Text Mining)
- Open Source
- Ease of Extensibility, Customisability (Python)
- Integrate not Re-implement
- "Web 2.0" – interactivity and dynamic interfaces

Cheshire3: Introduction

Today:

- Version 0.9.9 ... Almost ready for 1.0!
- Integrated:
  - Computational Grid for massively distributed processing
  - Data Grid for distributed petabyte scale storage
  - Data Mining for Classification, Clustering, Association Rule Mining
  - Text Mining for phrase extraction, natural language processing, semantic analysis
- Documentation and configuration applications almost finished
- Test and Example suite in progress

Cheshire3: Text Mining

Text Mining Tools Integrated:

- 4 part of Speech Taggers, including TreeTagger that supports multiple languages
- Phrase extraction library, plus additional custom implementation
- Deep Parser
- WordNet hierarchy
- Acronym recognition and resolution
- Gene/Protein recognition and resolution

- Excellent base toolkit... for modern English text.

Cheshire3: Computational Linguistics

- Computational Linguistics group at Liverpool using Cheshire3
- Analysis of Guardian newspaper articles

- Additional supported or improved features:
  - Ease of analysis of extracted text in index
  - Ease of collocation analysis (proximity word vectors)
  - Ease of extraction of sub-texts:
    - Spans between XML elements (eg 5 <lb/> tags)
    - Sentences/Paragraphs
  - Full proximity searching (word x within N words of word y)
  - Cross index proximity (word x within N of part-of-speech y)
  - SQL indexStores with term clustering
    (thanks to last year's HTM workshop!)

Text Mining for Historians

- For historians, Text Mining tends to focus on Patterns not Facts
- Modern History more prevalent, as more machine accessible documents available

- Classicists and Medievalists have fewer available tools, datasets
- Datasets have more issues (spelling, dates, text flow etc)
- NLP tools much harder to come by outside of Latin/Greek as language was very fluid

- So... other than the C.L. oriented patterns, what can we do?

- Enter some re-purposed data mining techniques...

Association Rule Mining

- Association Rule Mining (ARM) is a technique for discovering interesting patterns in transactions.
- Transactions in industry often shopping market trolleys
- Discovers patterns as rules:

  If someone buys a barbie doll, then they also buy a chocolate bar.
  If milk and bread and jam, then butter

- Can be used to increase profits (of course!) by putting the correct things on sale together, what to advertise together and assisting with shop floor layout.

- An example...

ARM Example

5 Shopping Baskets:

1: bread, butter, jam
2: bread, butter
3: bread, butter, milk
4: beer, bread
5: beer, milk

Simple statistics:

80% of baskets contain bread.
60% of baskets contain butter.
20% of baskets contain jam.

Associations:

100% of baskets that contain butter, also contain bread.
100% of baskets that contain jam, also contain butter.

ARM Example 2

5 Shopping Baskets:

1: bread, butter, jam
2: bread, butter
3: bread, butter, milk
4: beer, bread
5: beer, milk

Association Rules:

If butter, then bread

Support: 60% (60% of all transactions contain both)
Confidence: 100% (100% of butter transactions, also have bread)

If bread, then butter

Support 60%
Confidence 80% (Transaction 4 has bread, but not butter)

If jam, then bread and butter

Support 20% (Only 1 has jam)
Confidence 100% (Hence any rule is 100% confident)

ARM for Textual Analysis

Challenge:

Find arbitrary length groups of co-occurring words in different spans of text, without fixing any word(s) as required.

Text:

Early 15th C. Middle French MS describing the 100 years war.

Solution:

- Treat the words in a section of text as a basket:
  - 1 line span
  - 3 line span
  - 5 line span
  - Sentence
  - Column
  - Span between illuminated initials

Language Independent Textual Correlation Analysis

Method:

- Parse TEI XML transcription/edition of manuscript
- Split text based into lots of smaller records (baskets)
- Remove stopwords (de a le un par...)
- Build inverted indexes of the records, assigning identifiers
- Build vectors of which terms appear in which records

- Maybe do a search for a set of records to process
- Extract vectors from selected records, ignoring very frequent/rare
- Run Association Rule Mining application
- Turn numeric rules back into words
- Rank rules according to support

Language Independent Textual Correlation Analysis

Results for 5 lb span, stoplist, ordered by support:

[837]	roy france
[743]	roy engleterre
[484]	roy grant
[483]	messire jehan
[466]	gens armes
[400]	roy gens
[388]	roy messire
[381]	roy bien
[359]	grant gens
[317]	roy conte
[309]	roy englois
[298]	grant bien
[294]	messire conte
[287]	messire grant
[279]	roy duc
[275]	grant foison
[267]	messire gens
[263]	escuiers chevaliers
[253]	gens bien
[240]	grant france
[231]	france engleterre
[224]	messire bien
[223]	roy prince
[223]	roy pays

[219]	grant englois
[218]	messire chevaliers
[215]	messire france
[215]	sire messire
[210]	roy jehan
[205]	seigneurs roy
[196]	roy navarre
[193]	roy conseil
[191]	ville gens
[188]	englois bien
[187]	royaume france
[185]	grant conte
[184]	grant engleterre
[184]	roy france engleterre
[184]	roy chevaliers
[182]	ville grant
[180]	france duc
[180]	roy devant
[177]	gens france
[176]	messire duc
[176]	grant armes
[175]	grant chevaliers
[173]	grant devant
[172]	royaume roy

[172]	roy phelippe
[171]	gens englois
[170]	roy comment
[169]	robert messire
[167]	france conte
[166]	normandie duc
[166]	temps roy
[165]	gens conte
[162]	messire armes
[162]	gens devant
[161]	saint roy
[161]	roy armes
[160]	ville roy
[160]	pays grant
[160]	pays gens
[159]	messire engleterre
[158]	monseigneur jehan
[158]	bien armes
[157]	temps grant
[157]	prince galles
[156]	messire englois
[156]	roy fait
[155]	jehan conte
[154]	duc conte

Language Independent Textual Correlation Analysis

Discovers frequent patterns:

- Top rule: 'king' and 'france' appear together a lot because the text is about who is the rightful King of France.
- Titles appear together ('messire', 'sire', 'monseigneur', 'conte', 'duc', 'prince', 'roy')
- Descriptive terms for people appear together ('chevalier', 'escuier', 'gens [d'armes]') along with titles

But our definition said we wanted **interesting** patterns.

These are interesting in as much as they show a good set of topics for the text, but they're immediately obvious given an understanding of the text.

Need a new metric to order the patterns by.

Language Independent Textual Correlation Analysis

Order by 'surprise':

- Determine the frequency of the terms in the rule, and find the likelihood that the terms will appear together at random (eg bien grant)
- Order by largest proportional difference between expected and actual number of co-occurences
- Favours longer sets of terms, possibly with low support
- Also discovers negatively correlated terms – sets of terms that appear together less frequently than would be expected given their overall frequency

Much more interesting results...

Language Independent Textual Correlation Analysis

Results, 5 lb span, stoplist, ordered by 'surprise':

Expected (%)	Actual (%)	Word1	Word2	Word3	Word4
[0.006 (0000%)]	[24 (02%)]	mil (80)	ccc (39)	an (185)	
[0.015 (0000%)]	[33 (03%)]	mil (80)	iii.c (104)	an (185)	
[0.013 (0000%)]	[27 (03%)]	villes (206)	citez (62)	chasteaux (104)	
[0.016 (0000%)]	[23 (02%)]	mil (80)	grace (110)	an (185)	
[0.022 (0000%)]	[24 (02%)]	mainte (77)	armes (788)	appertise (37)	
[0.077 (0001%)]	[34 (03%)]	messire (2710)	guichart (56)	angle (51)	
[0.071 (0001%)]	[25 (02%)]	sicomme (193)	oy (195)	avez (190)	
[0.071 (0001%)]	[24 (02%)]	oy (195)	cy (193)	avez (190)	
[0.090 (0001%)]	[25 (02%)]	roy (3677)	messire (2710)	claquin (83)	bertrain (110)
[0.086 (0001%)]	[22 (02%)]	saint (583)	mil (80)	an (185)	
[0.121 (0001%)]	[31 (03%)]	messire (2710)	eustace (107)	aubrechicourt (42)	
[0.246 (0002%)]	[58 (06%)]	messire (2710)	claquin (83)	bertrain (110)	
[0.127 (0001%)]	[30 (03%)]	fer (85)	bien (1675)	armeures (90)	
[0.128 (0001%)]	[26 (03%)]	messire (2710)	harecourt (78)	godefroy (61)	
[0.143 (0001%)]	[26 (03%)]	messire (2710)	captal (113)	beus (47)	
[0.164 (0002%)]	[27 (03%)]	monseigneur	charles (268)	blois (79)	
[0.285 (0003%)]	[46 (05%)]	guichart (56)	angle (51)		
[0.166 (0002%)]	[26 (03%)]	prevost (49)	marchans (34)		
[0.166 (0002%)]	[26 (03%)]	oy (195)	comment (452)	avez (190)	
[0.135 (0001%)]	[21 (02%)]	thomas (152)	messire (2710)	felleton (33)	
[0.185 (0002%)]	[28 (03%)]	monseigneur	charles (268)	bloys (89)	
[0.192 (0002%)]	[29 (03%)]	monseigneur	mauny (150)	gautier (165)	
[0.192 (0002%)]	[24 (02%)]	grant (2127)	foison (309)	escuiers (321)	chevaliers (918)
[0.230 (0002%)]	[28 (03%)]	robert (297)	monseigneur	artois (100)	
[0.241 (0002%)]	[28 (03%)]	dames (78)	damoiselles (31)		
[0.337 (0003%)]	[38 (04%)]	contes (105)	barons (352)	chevaliers (918)	
[0.246 (0002%)]	[25 (02%)]	pennons (32)	banieres (77)		

Language Independent Textual Correlation Analysis

Ordered by support useful to see common correlations

- Tends to discover common pairings, related to the overall topic.

Ordered by surprise useful to see 'interesting' correlations.

- Tends to discover 'tag line' phrases, names, and sub-topical correlations.

Other orderings possible, but not investigated yet.

Other than trivial stopword list, no language dependencies.

- Future research will investigate on historical text where language models are available.

Synergies with existing methods

- Provides information as to where to start investigating in more detail with Computational Linguistic techniques

Thank You!

Questions?