‘Milk, bread and toothpaste’: Adapting Data Mining techniques for the analysis of collocation at varying levels of discourse

Rob Sanderson, Matthew Brook O’Donnell and Clare Llewellyn
What happens with all that supermarket loyalty card data?
What happens with all that supermarket loyalty card data?
What happens with all that supermarket loyalty card data?

- Bread & Milk: 4/6 (67%)
  - Bread -> Milk: 80% (4/5)
  - Milk -> Bread: 100% (4/4)
What happens with all that supermarket loyalty card data?

- Bread & toothpaste: 2/6 (34%)
- Bread → toothpaste: 40% (2/5)
- Toothpaste → bread: 100% (2/2)
Corpus linguistic context: Collocation

• Sinclair – discoverer of collocation
• focus on pairs – node & (significant) collocate
• 4 word optimum window size (OSTI etc.)
• statistics well understood (Evert, Church etc.)
• extensions/developments
  • filtering (per sentence, POS, constituents) – (Evert)
  • targeted analysis – collostruction measures (Gries & Stefanowitch)
  • levels below and above the word (Baayen)
  • search for ngrams, frames and sets of more than two words
  • Concgrams (Cheng, Greaves & Warren)
What can ARM offer to corpus linguistics?

- base level – find collocate pairs
- reduce directionality & node focus
- find sets of 2+ co-occurring items
- find associations beyond the 9 word window at various levels of discourse
ARM overview: Terminology

- ITEM SETS
- ASSOCIATION RULES
- SUPPORT
- CONFIDENCE
Association Rule Mining for Text

- Requires Integers not strings -- one per type
- No Order or Distance
- No Frequency (but some algorithms have weight)
- Apply to a range of discourse levels
  - Window of N words (sliding, discrete, centered on term)
  - Split at verb forms
  - Clause (as per shallow parser)
  - Sentence
  - Paragraph [subsection, section, chapter]
  - Full Text
Desirables for Text ARM

- No stoplist (or very minimal)
- No requirement for PoS tagging
- Flexible ‘unit’ (transaction) sizes
- Not impossible amounts of data to look through
- Order of results is important [cf google]
Initial Experiments

• Using 100,000 articles from the Home News section of *The Guardian* 1998-2004 (~45 mill. wds)

1. 11 word span around a selected node within sentence & whole sentence search

2. unfocused rule generation with an iterative mining process

3. mining rules across an entire article from articles containing words from CONTROVERSY set
Example 1: the fact that

Procedure
a. using a flexible 11 word window, stopping at sentence breaks
b. retrieving whole sentences containing the fact that

• stoplist: a, an, the, and, fact, that
• support threshold: 0.1% (i.e. item found in 5+ sentences)
• confidence threshold: 30% (A --> B, B with A in 30+% of A instances)

Results: Summary
• 5971 windows containing the fact that in 4198 records
• 4480 sentences containing the fact that

• 6253 item sets found
• 12868 rules
### Example 1: ‘the fact that’ -- item sets

<table>
<thead>
<tr>
<th>waking to</th>
<th>to ignore</th>
<th>of in</th>
</tr>
</thead>
<tbody>
<tr>
<td>woken to</td>
<td>praised introducing</td>
<td>drew to</td>
</tr>
<tr>
<td>of regardless</td>
<td>of highlighted</td>
<td>to reasonable</td>
</tr>
<tr>
<td>of reminded</td>
<td>reflected in</td>
<td>of lies</td>
</tr>
<tr>
<td>of ignore</td>
<td>to resigned</td>
<td>to reflect</td>
</tr>
<tr>
<td>of spite</td>
<td>hide to</td>
<td>lies in</td>
</tr>
<tr>
<td>of cannabis</td>
<td>of combined</td>
<td>reference to</td>
</tr>
<tr>
<td>of underlined</td>
<td>of proud</td>
<td>to respond</td>
</tr>
<tr>
<td>spite in</td>
<td>of sight</td>
<td>of proportion</td>
</tr>
<tr>
<td>alerted to</td>
<td>to referring</td>
<td>of to</td>
</tr>
<tr>
<td>stated in</td>
<td>lie in</td>
<td>glasgow in</td>
</tr>
<tr>
<td>comfort in</td>
<td>sharp in</td>
<td>to in</td>
</tr>
</tbody>
</table>

*for 11 word window search*
Example 1: ‘the fact that’ -- rules

the fact that

- complicated --> by
- spite --> in
- compounded --> by
- may been --> have
- trying --> to
- according --> to
- referring --> to
- think there --> i
- secret i --> no of
- heightened --> by
- wake --> up to
- stems --> from
- stated --> in
- referred --> to
- much made --> of

for 11 word window search

all rules have 100% confidence
Example 1: ‘the fact that’ -- rules

- due -> to
- trying -> to
- led -> to
- spite -> in of
- unable -> to
- none -> of
- attempt -> to
- according -> to
- kind -> of
- sort -> of
- number has -> of
- number been -> of
- proud -> of

for sentence search
Example 1: trying to + the fact that

the fact that X trying to Y

(1) But the fact that someone is trying to sell these weapons in the US is nothing new.

(2) ... and the fact that the department is trying to disown talk of targets at this late stage tells its own story.

(3) The fact that they are trying to ride the argument of inevitability demonstrates actually how weak their case is ...
Example 1: trying to + the fact that

trying to Z the fact that

(4) ... Thomas White, accused the Pentagon of trying to gloss over the fact that troops would remain for months.

(5) I am just trying to come to terms with the fact that I have seen a dead body for the first time in 10 years.

(6) Mr Riaz said his party was trying to address the fact that each time a white member of the party mentioned asylum or immigration, it was interpreted as racist.
Example 2: sentence level rules

- Uses every first sentence (TISC) from 100,000 articles in lemmatized form

- Stoplist: *the a an and be*

- Strongest correlations appeared to be phrases e.g. *Sinn Fein, Tony Blair, Prime Minister*

- Implemented merging procedure to capture these and treat them as single terms in a second iteration
Example 2: text-initial sentences

<table>
<thead>
<tr>
<th>w1</th>
<th>w2</th>
<th>adj.</th>
<th>freq</th>
<th>%</th>
<th>ordering</th>
</tr>
</thead>
<tbody>
<tr>
<td>fein</td>
<td>sinn</td>
<td>265</td>
<td>265</td>
<td>100%</td>
<td>'sinn fein’ 265</td>
</tr>
<tr>
<td>alan</td>
<td>milburn</td>
<td>206</td>
<td>206</td>
<td>100%</td>
<td>'alan milburn’ 206</td>
</tr>
<tr>
<td>handling</td>
<td>of</td>
<td>160</td>
<td>162</td>
<td>98%</td>
<td>'handling of’ 160</td>
</tr>
<tr>
<td>accord</td>
<td>to</td>
<td>3443</td>
<td>3462</td>
<td>99%</td>
<td>'accord to’ 3445</td>
</tr>
<tr>
<td>cent</td>
<td>per</td>
<td>152</td>
<td>152</td>
<td>100%</td>
<td>'per cent’ 167</td>
</tr>
<tr>
<td>ministry</td>
<td>of</td>
<td>484</td>
<td>519</td>
<td>93%</td>
<td>'ministry of’ 485</td>
</tr>
<tr>
<td>of</td>
<td>string</td>
<td>144</td>
<td>163</td>
<td>88%</td>
<td>'string of’ 142</td>
</tr>
<tr>
<td>to</td>
<td>try</td>
<td>1256</td>
<td>1369</td>
<td>91%</td>
<td>'try to’ 1220</td>
</tr>
<tr>
<td>for</td>
<td>responsible</td>
<td>301</td>
<td>318</td>
<td>94%</td>
<td>'responsible for’ 302</td>
</tr>
<tr>
<td>accuse</td>
<td>of</td>
<td>1094</td>
<td>2082</td>
<td>52%</td>
<td>-</td>
</tr>
<tr>
<td>connection</td>
<td>in with</td>
<td>404</td>
<td>416</td>
<td>97%</td>
<td>'in connection with’ 404</td>
</tr>
</tbody>
</table>
Example 2: text-initial sentences – item sets

- foot
- mouth
- chancellor
- gordon_brown
- student
- university
- united
- state
- pupil
- school
- right
- human
- nhs
- health
- secretary
- david_blunkett
- foreign
- office
- hospital
- patient
- teacher
- school
- union
- european
- nhs
- hospital
- general
- election
- england
- wales

- education
- school
- there
- no
- cabinet
- minister
- accord_to
- survey
- child
- parent
- union
- leader
- more
- than
- find
- body
- conservative
- party
- party_tory
- david_blunkett
- home
- service
- health
- war
- iraq
- london
- ken_livingstone
- court
- judge
- mayor
- london
### Example 2: text-initial sentences – rules

<table>
<thead>
<tr>
<th>Phrase</th>
<th>Replacement</th>
<th>Target Word</th>
<th>Probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>accuse after</td>
<td>--&gt; of</td>
<td>of</td>
<td>97%</td>
</tr>
<tr>
<td>accuse by</td>
<td>--&gt; of</td>
<td>of</td>
<td>97%</td>
</tr>
<tr>
<td>allow for</td>
<td>--&gt; to</td>
<td>to</td>
<td>91%</td>
</tr>
<tr>
<td>urge</td>
<td>--&gt; to</td>
<td>to</td>
<td>89%</td>
</tr>
<tr>
<td>time first</td>
<td>--&gt; for</td>
<td>for</td>
<td>87%</td>
</tr>
<tr>
<td>decision</td>
<td>--&gt; to</td>
<td>to</td>
<td>86%</td>
</tr>
<tr>
<td>plan on</td>
<td>--&gt; to</td>
<td>to</td>
<td>85%</td>
</tr>
<tr>
<td>part</td>
<td>--&gt; in</td>
<td>in</td>
<td>84%</td>
</tr>
<tr>
<td>end at</td>
<td>--&gt; of</td>
<td>of</td>
<td>84%</td>
</tr>
<tr>
<td>plan government</td>
<td>--&gt; to</td>
<td>to</td>
<td>84%</td>
</tr>
<tr>
<td>body find</td>
<td>--&gt; of</td>
<td>of</td>
<td>84%</td>
</tr>
<tr>
<td>pledge</td>
<td>--&gt; to</td>
<td>to</td>
<td>83%</td>
</tr>
<tr>
<td>give government</td>
<td>--&gt; to</td>
<td>to</td>
<td>83%</td>
</tr>
<tr>
<td>plan his</td>
<td>--&gt; to</td>
<td>to</td>
<td>83%</td>
</tr>
<tr>
<td>victim</td>
<td>--&gt; of</td>
<td>of</td>
<td>83%</td>
</tr>
<tr>
<td>convict</td>
<td>--&gt; of</td>
<td>of</td>
<td>82%</td>
</tr>
<tr>
<td>describe</td>
<td>--&gt; as</td>
<td>as</td>
<td>81%</td>
</tr>
<tr>
<td>plan new</td>
<td>--&gt; to</td>
<td>to</td>
<td>81%</td>
</tr>
<tr>
<td>murder on</td>
<td>--&gt; of</td>
<td>of</td>
<td>81%</td>
</tr>
</tbody>
</table>
Example 3: ARM at the article level

- Retrieve articles containing at least one instance of \{controversy, row, embarrassment, blow\} in its first sentence (TISC)
- POS filter for verbs, nouns, adjectives & adverbs
- Stoplist: be, have
Example 3: ARM at the article level

- resign
- remark
- blow
- complaint
- defend
- adviser
- dismiss
- relation
- row
- leadership
- criticism
- shadow
- document
- smith
- protest
- declare
- oppose
- supporter
- radio
- criticise
- chancellor
- allegation
- downing
- investigate
- william
- urge
- conduct
- agreement
- discuss
- sunday
- june
- opposition
- regard
- reject
- spokeswoman
- brown
- civil
- threaten
- ban
- doubt
- current
Example 3: ARM at the article level

2 item sets
downing_street prime_minister
conservative tory
mr_blair prime_minister
office row
issue row
labour downing_street
secretary row
minister row
secretary shadow
tony_blair prime_minister
labour cabinet
comment page
downing_street minister
row claim

3 item sets
conservative party tory
labour mp tory
mr_blair prime_minister when
labour party tory
labour election party
labour mp party
one about row
labour mp minister
when make row
make about row
take make row
when about row
or about row
over make row
over when row
over when row
Summary

- association rule mining is a useful tool to add to the corpus linguistic methodology
- appears to reduce the effect of noise allowing the use of much wider windows, even up to the whole text level
- has potential applications in
  - phraseology
  - ‘pattern grammar’/construction discovery
  - complex word associations (similar to concgrams)
- need for refinements and wider evaluation (other corpora and comparison with range of collocational measures)