The Computable News Project
Milestone 3
Who is talking about what?

James Curran

Schwa-lab
School of Information Technologies
The University of Sydney

Capital Markets CRC Limited

Computable News team

- Postdoctoral fellows:
  - **Ben Hachey**
  - **Matthew Honnibal**
  - **David Vadas**
- Project manager:
  - **Tim Yeates**
- PhD students:
  - **Joel Nothman**
  - **Will Radford**
  - **Tim O'Keefe**
- Design consultant:
  - **Will Cannings**
- additional annotators from Schwa-lab and outside

Capital Markets CRC Limited

M3 Deliverables

- ✔ Demo: who is talking about what?
- ✔ Quotation extraction and attribution tools
- ✔✔ Quotation annotation tool
- ✔✔ 800 SMH articles annotated with quotations
- ✔ Finalised event annotation scheme and tool design
- ✔ ~~Sentiment of quotations and attitudes across time~~
- ✔✔ Web-sourced quotation annotation using Freelancer

Capital Markets CRC Limited

M3 Demo

- ✔ Dynamic entity/topic pages from Sydney Morning Herald data
  - Photos view: photos for entities/topics
  - Quotes timeline view: Quotations attributed to entities
  - Discussion timeline view: Quotes, tweets about entities/topics
  - Topics view: Related topics for entities/topics
  - Timeline view: Stories, quotes, etc. for entities/topics
  - Updated story view featuring quotations
- Detection and tracking of non-entity topics of interest
- Outsourced quotation annotation from the web using Freelancer
  - Collected largest quote corpus ever for under $700

Capital Markets CRC Limited

M3 Demo: Feature checklist

- ✔ lists of quotations by or about entities
- ✘ lists of quotations about events
- ✘ timeline when statements are quoted elsewhere (e.g., on Twitter, blogs, Wikipedia, or other news sources)
- ✘ graphs that visualize the connections between speakers and the entities/events they talk about
- ✘ ~~aggregated sentiment expressed by or about given entities~~

Capital Markets CRC Limited

M3 Demo: Bonus features

- ✔✔ automatic topic detection and tracking
- ✔✔ dynamic topic pages
- ✔✔ improved story view with quotations
- ✔ redesigned database for efficiency and Fairfax Entity Store

Capital Markets CRC Limited

What is a topic?

The opposition spokesman on climate action, Greg Hunt, said: ‘‘This issue will be resolved well before 2016. If the Coalition is elected on the basis of scrapping the carbon tax, Labor must support its removal, including voting for its abolition in the House of Representatives and the Senate.’’

Based on http://www.theage.com.au/environment/climate-change/ending-carbon-tax-may-be-long-road-for-coalition-20110821-1j4sn.html

Capital Markets CRC Limited

What is a topic?

> The opposition spokesman on climate action, Greg Hunt, said: ‘‘This issue will be resolved well before 2016. If the Coalition is elected on the basis of scrapping the carbon tax, Labor must support its removal, including voting for its abolition in the House of Representatives and the Senate.’’

- carbon tax $\rightarrow$ `wiki/Carbon_tax`

Capital Markets CRC Limited

What is a quotation?

Failed entrepreneur and former CEO Eddy Groves will defend a criminal charge relating to the collapse of Australia's biggest childcare chain, ABC. ''I've pleaded not guilty today and I will vigorously defend the charge, and that's really all I can say right now,'' Groves said.

Based on http://www.news.com.au/business/breaking-news/abc-learning-founder-eddy-groves-pleads-not-guilty/story-e6frfkur-1225996040971

Capital Markets CRC Limited

Who is talking?

Failed entrepreneur and former CEO Eddy Groves will defend a criminal charge relating to the collapse of Australia's biggest childcare chain, ABC.
''I've pleaded not guilty today and I will vigorously defend the charge, and that's really all I can say right now,'' Groves said.

- **Quote**: I've pleaded not guilty today and I will vigorously defend the charge, and that's really all I can say right now
- **Speaker**: Eddy Groves (wiki/Eddy_Groves)

Capital Markets CRC Limited

About what?

The opposition spokesman on climate action, Greg Hunt, said: ‘‘This issue will be resolved well before 2016. If the Coalition is elected on the basis of scrapping the carbon tax, Labor must support its removal, including voting for its abolition in the House of Representatives and the Senate.’’

Capital Markets CRC Limited

Spokespeople

- Many quotes are delivered by spokespeople who are not themselves of public interest.
- When annotating the documents, we therefore instructed annotators to mark speakers' affiliations.
- `Opposition spokesman on climate action, Greg Hunt`
- Greg Hunt → The opposition → Liberal Party of Australia
- Relation extraction is difficult, but the terms spokesman, spokesperson, etc are very useful clues.

Capital Markets CRC Limited

Cost-effective Freelance annotation

- We need manually annotated example data to build and evaluate our tools.
- Milestone 1 data was annotated by project members, PhD students, and research assistants.
- In Milestone 2 we took advantage of our web-based annotation tool by using Freelancer.com to hire annotators.
- This Milestone, Freelancer.com allowed us to collect **400,000 words of quotation annotation for $700**.
- Most Milestone 2 annotators bid for work on Milestone 3.
- Annotators were paid an average of $7/h.
- Most annotators were from the US.

Capital Markets CRC Limited

Quote results

- Quotation extraction is 99.34% accurate, with the few errors coming from missing quotation marks
- Quotation attribution performance is highly dependent on the accuracy of the linker:

System	Precision	Recall	F-score
Milestone 2 linker	25.45%	23.85%	24.63%
Gold standard	91.28%	91.26%	91.27%

Capital Markets CRC Limited

M3 Quotes: deliverables

- ✔ Tools for quotation extraction and attribution
- ✔ Tools for identifying quotations about given entities/topics
- ✔ Quotation lists and timelines for a given entity
- ✘ Connection graphs for speakers and the entities they talk about (not displayed, but we have the data)

Capital Markets CRC Limited

M3 Quotes: bonus

- ✔✔ Quotation annotation tool
- ✔✔ 800 SMH articles annotated with quotes

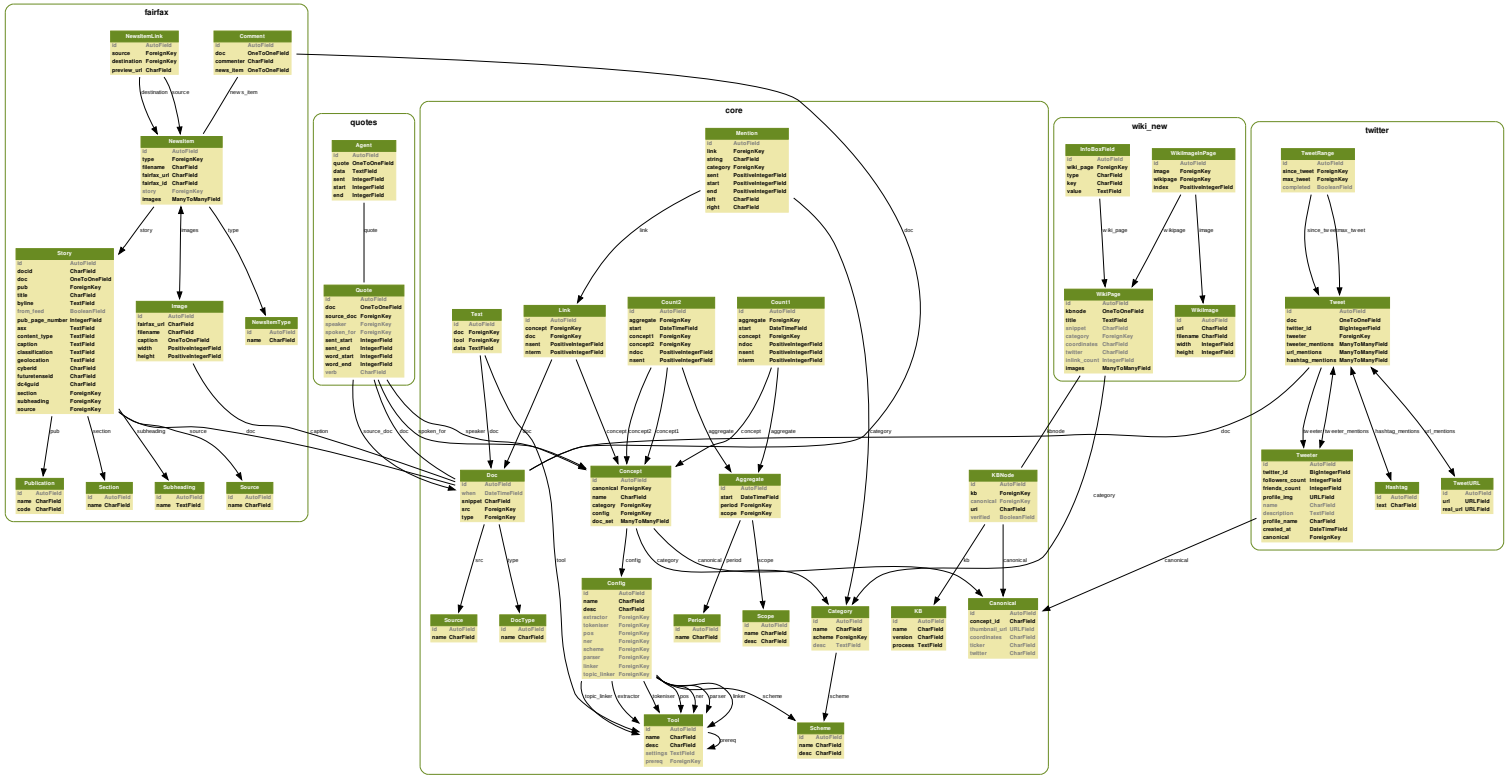
Capital Markets CRC Limited

DB refactor

- Single `Doc` model to represent content of all text documents (e.g., stories, tweets, quotes)
  - `Text` objects represent various levels of processing from NLP stack
  - `Link` objects represent entity/topic mentions
  - `Config` represents NLP tool versions used to extract mentions
- Introduction of `KBNode` and `Canonical` tables, mapping
  - from `Concepts` in given configurations
  - to external resources (e.g., future entity store API)

Capital Markets CRC Limited

New schema



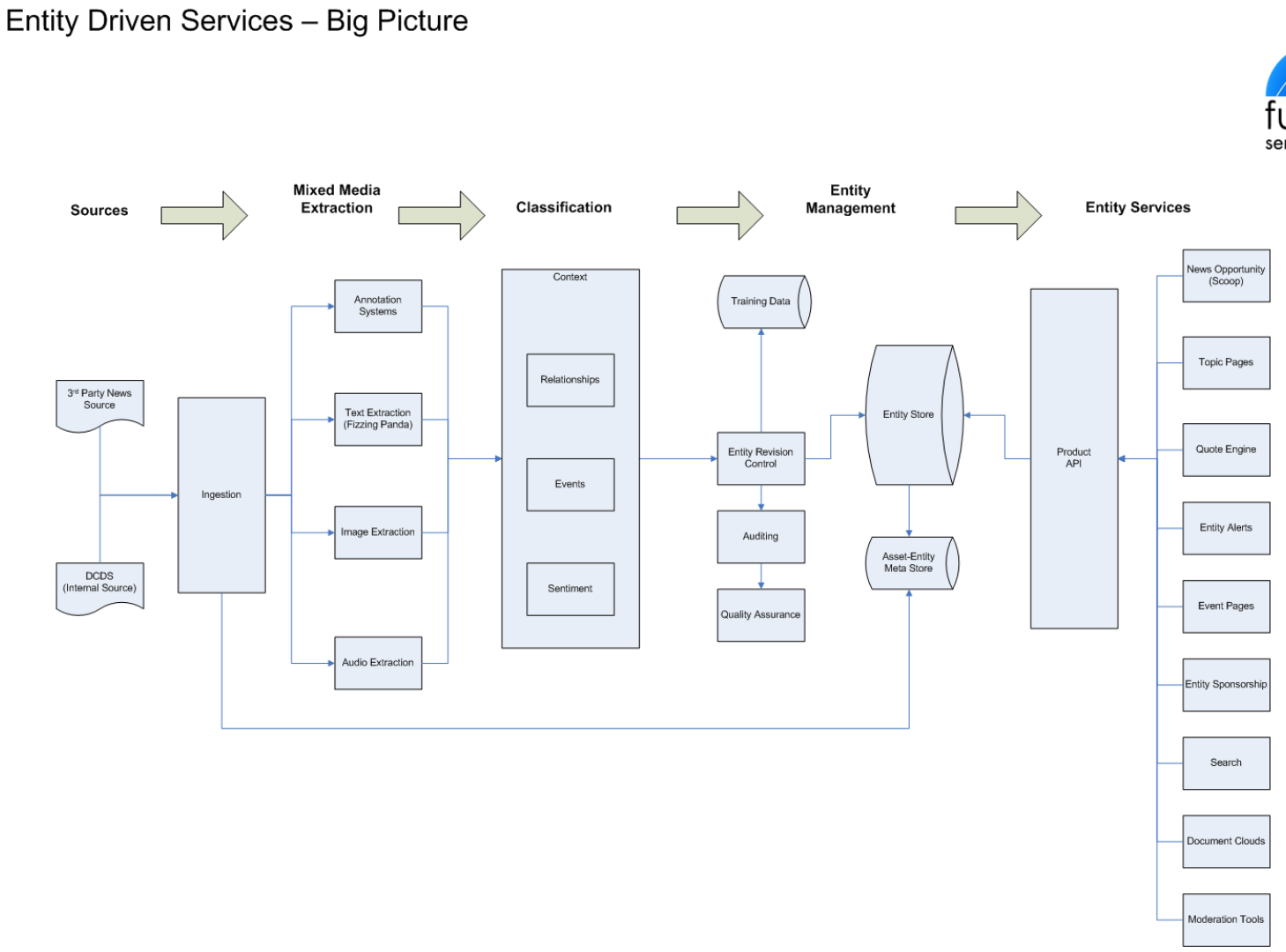
Capital Markets CRC Limited

DB optimisation and caching

- New `Count1`, `Count2` cache tables store aggregates
- More indices on tables, Django efficiency improvements
- Using memcached to cache trending and concept pages.

Capital Markets CRC Limited

Entity store plan (George Wright)



M3 Tech: bonus

- ✔✔ Database refactor, in preparation for entity store API
- ✔✔ Database optimisation and caching

Capital Markets CRC Limited

Natural language processing stack

- **Tokenisation**
- Part-of-speech tagging
- Named entity recognition
- **Coreference resolution**
- **Entity linking**
- Topic linking

Left quotes in Fairfax stories, let me count the ways. . .

‘	`
&apos;	&39;
‘ (U+2018)	‘ (U+201b) [1]
&#8216;	&#x2018;
&lsquo;	0x91
0xe2 0x80 0x98	0xe2 0x80 0x9b
&#8219;	&#x201b;
0xe2 0x9d 0x9b	&#10075;
&#x275b;	

[1] Couldn't render this properly in LaTeX

Capital Markets CRC Limited

Exploring in-document co-reference

- We use naïve heuristics to map entity mentions within a document.
  - `Mr Howard` $\rightarrow$ `John Howard`
  - `Howard` $\rightarrow$ `John Howard`
- But what about the last 20%?
  - `Mrs Howard` $\rightarrow$ `?`
  - `Jack Howard` $\rightarrow$ `?`

Capital Markets CRC Limited

Evaluating entity linking

- More demanding evaluation over the COMPNEWS annotated data - how well can we link **each** entity mention?

System	Precision	Recall	F-score
Milne-based	32.57	30.29	31.38
Cucerzan-based	63.55	59.10	**61.25**

- Built error analysis tools.
- Significant engineering effort to reduce linking runtime of 730 test stories from 75 mins to 15 mins.

Capital Markets CRC Limited

Academic progress

- Revisions to Artificial Intelligence Journal paper resubmitted for comment.
- Paper accepted to Web Information System Engineering 2011 conference.
- Participation in Text Analysis Conference Entity Linking shared task 2011.

Capital Markets CRC Limited

M3 NLP: bonus

- Improved tokenisation
- Improved coreference resolution
- Improved entity linking

Event Linking example

Mr Dutton won Dickson from Labor's Cheryl Kernot in 2001. Ms Kernot won the seat for Labor in 1998 after defecting from the Democrats. On the night of the 1998 election, with the result close and yet to be finalised, Ms Kernot lost her cool on national television, thinking she had lost. She berated Labor for not finding her a safe seat.

1. won: Kernot Takes a Pounding
2. won: Opponents join to take shine off Kernot's win
3. defecting: Kernot's Labor gamble
4. election: Election over, but the battle has just begun
5. lost, berated: Outburst by Kernot 'intemperate'

Capital Markets CRC Limited

Event Linking Scheme (1)

1. Find an event-denoting expression
2. Ignore it if:
   - hypothetical or uncertain
   - Not newsworthy, including:
   - Wrong semantic class (reporting, perception, etc.)
   - Non-newsworthy occurrence
3. Otherwise:
   - Select a single word expressing the event
   - If you have already marked another mention for the same event (or a closely related event first reported in the same article):
     - If that mention is in the same sentence, ignore it.
     - If that mention is in another sentence, mark the new mention as part of the same event.
   - Otherwise, mark the new mention as a new event.

Capital Markets CRC Limited

Event Linking Scheme (2)

4. Select a category for the event:
   - Basic event – probably first reported in one news article
   - Complex event – likely to have multiple articles; often a named event
   - Trend or measured change
   - Many specific events
   - Non-specific
5. If a basic event:
   - Try to link to the article first reporting that event as having happened, or:
     - First reported here
     - Precedes 1986
     - Not found, which includes: No mention in archive, Not reported in archive, Not reported after occurrence
6. If a complex event:
   - Try to link to a Wikipedia article specifically about the event
   - Or mark as Not found.

Capital Markets CRC Limited

Annotation costs

Employee type	Price per hour	Hours spent	Net price
Post-docs	$50	112	$5,600
'Volunteer' PhD students	$35	54	$1,890
Project PhD students	$40	51	$2,040
Research assistant	$35	51	$1,785
Milestone 1 total	$42.2	348	**$11,315**
Milestone 2 total	$10	250	**$2,500**
Milestone 3 total	$7	100	**$700**

Capital Markets CRC Limited