The Computable News Project
Milestone 7
Production NEL

James Curran

ə-lab
School of Information Technologies
The University of Sydney

Capital Markets
CRC Limited

Computable News team

- Project leader: **James Curran**
- Project manager: **Candice Loxley**
- Postdoctoral researchers:
  - **Daniel Tse**
  - second position still being filled (2 year position)
- Senior Software Developer: **Will Cannings**
- PhD students:
  - **Joel Nothman**
  - **Will Radford**
  - **Tim O'Keefe**

Capital Markets CRC Limited

Computable News team

- PhD students on related work:
  - **Tim Dawborn**
  - **Andrew Naoum**
  - **Glen Pink**
- Summer Scholars:
  - **Fergus Macpherson**
  - **Sebastian Pauka**
  - **Kristy Hughes**

M7: Engineering

- ✔✔ Supported high-profile Melbourne Cup demo
- ✔✔ On site support for My Masthead development environment
- ✔✔ Code cleanup, installation and improved test-case coverage
- ✔✔ Finalised database testing to enable Cassandra decision
- ✔✔ Switched NEL backend to Cassandra
- ✔✔ Extensive consultation about NEL and Cassandra optimisation
- ✔✔ Large-scale, fast NEL with Cassandra on over 150 execution hosts

M7: NEL

- ✔✔ TAC 2012 full results
  - ✔ Improvement from 72.44 F1 to 73.96
    - More data cleaning
    - Local context features

M7: Facts & Opinions

- ✔ Preliminary apposition extraction
- ✔ Preliminary slot filling
- ✔ Preliminary numeric fact extraction
- ✔ Refined quote-opinion annotation scheme
- ✔ All quotes have at least been double-annotated with opinions
- ✔ Quotes system working in docrep

Capital Markets CRC Limited

M7: Events

- ✔ Built a corpus of hyperlinks within DCDS data
- ✔ Annotated a subcorpus for event links
- ✔ Built a basic classifier to replicate this annotation
- ✔ Zone weighting for hyperlink detection $\Rightarrow$ 17% relative gain

M7: Summer Scholarship Programme

- ✔ Hired three exceptional students
- ✔ All working on key CompNews problems
- ✔ 12 weeks of engaging research work
  - gives the students a taste of research
  - hopefully gives the students a taste **for** research
  - gives PhD students supervision experience

Capital Markets CRC Limited

Melbourne Cup demo

- Ported demo timeline to `timeline.js`
- Integrated new design
- Implemented link scoring to counter crowded timelines
- Prototype tools for curating timelines
- Linked and loaded racing stories
- New type of entity: HORSE

Capital Markets CRC Limited

Extensive support

- Spent after hours time on site to assist in deploying FP for My Masthead
- As a result of this, we packaged required linking data and created automated installation scripts to simplify deployment
- Contributed to database testing, performance testing, product discussions and code workshops

Capital Markets CRC Limited

Software Engineering

- Trimmed dependencies
- Full-use of `python` dependency management tools:
  - `pip`
  - `easy_install`
- New prompt-less installation scripts can be used to install FP with no interaction
- Set up VM images simulating the Fairfax architecture for testing

Capital Markets CRC Limited

Database testing conclusions

- Completed comprehensive read and write testing of Cassandra and Hypertable
- Hypertable was >2x faster at reads, >4x faster at writes
- Because of ops familiarity, Cassandra was selected
- Current low read/write requirements means Cassandra performs acceptably
- Performance will need to be reviewed when counts 2 data is integrated

Capital Markets CRC Limited

Faster linking

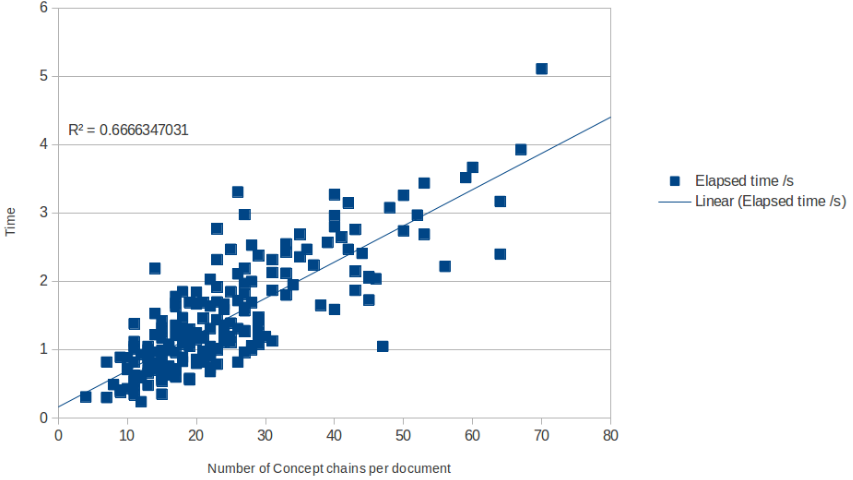
- Changing databases presented an opportunity to refactor the linking code
- Linking data is now stored in an optimised, pre-computed format
- Less data is transferred, and the linkers themselves are simplified
- This is less flexible than computing data as needed, but improves linking speed
- Average per-doc speed on a small document set (SMH 2012-01-01) is 1.06s
- Overhead will mean documents linked through the API will take longer than this, but will on average take less than 5s

Capital Markets CRC Limited

Batch linking

- During university holidays 157 undergraduate desktop machines are available as a cluster
- A bulk linking queue system was created to utilise this
- Multiple Cassandra and Solr replicas were set up to distribute load
- 1 year of articles (Dec 2011 - Nov 2012) from 4 mastheads (SMH, The Age, WA Today and Brisbane Times) were used to test performance
- Linking took 1.95hrs (avg 4.5s per doc on each node), and produced 15.8gb of data

Number of concepts determines document linking speed



tal
ets
CRC Limited

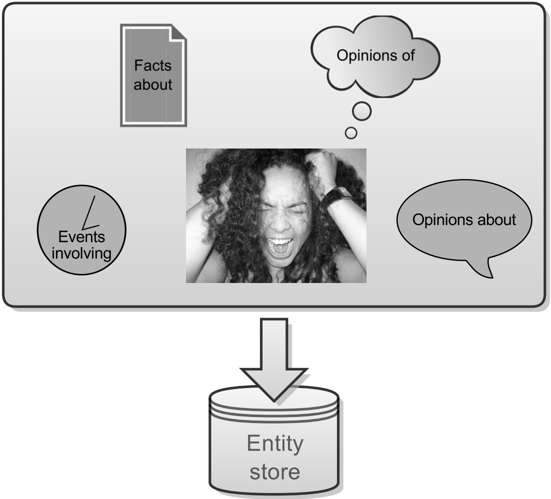
Consolidating TAC 2012 success

- Wrote and submitted system description paper
- Early access to other system description papers
- We are ranked #2 - 3.1% off the pace for KB linking

Team	KB B^{3+}
Microsoft Research	68.7
CMCRC	65.6

Capital Markets CRC Limited

Characterising entities



What do we know about our entities?

- News includes facts about entities to give the reader context
  - age
  - job description
  - spouse
  - stock ticker
  - ...
- We want to extract facts and use them for
  - improving linking accuracy
  - building a knowledge base
  - automatically enriching news

Apposition, a structure for adding information, . . .

- Readers need to understand a story's context, but some entities require extra work on behalf of the journalist

Ambiguous entities

Hey Dad! star **Robert Hughes** will plead not guilty . . .

Novel entities

His lawyer, **Greg Walsh**, says . . .

- Apposition relates a **head** Noun Phrase with one or more coreferent *attribute* Noun Phrases
- Punctuation is a reasonable cue, but not a silver bullet

Capital Markets CRC Limited

Apposition for Named Entity Linking

- Precise, local and strong attributes are well-suited to NEL
- Can apposition help
  - disambiguate known concepts?
  - disambiguate unknown concepts (i.e., two "John Smiths")?
  - extract enough textual information for fast, database NEL?

```
SELECT concept
WHERE name = 'Robert Hughes'
AND role = 'actor'
```

Finding person appositions

- Using sentence context (a rough approximation) increases performance from 71.41 to 73.96
- We have simple patterns over POS and NER tags.

HEAD , ATTR ,

Ntaras, *a trained cage fighter*, ...

ATTR , HEAD ,

The victim, 19-year-old **Nicholas Barsoum**, ...

- We plan to incorporate apposition detection into NEL

Extracting facts for Slot Filling

- Apposition is highly specific and is narrow coverage
- Slot filling involves extracting named facts about specific entities
- This requires identifying
  - Entities (i.e., NEL)
  - Values (this can require inference)
  - Value types (including normalisation)
  - Slots those values fill
  - Resolving duplicate or conflicting slot values

Slot filling

Place and date of birth

Robert Hughes was *born* in Sydney on July 28, 1938 into a prominent family of lawyers and politicians.

Age, knowing that the article was published 21/11/2012

. . . signed an extradition order for Hughes after a London magistrate in September determined that the 64-*year-old* return to NSW for questioning.

Slot filling

Number of employees or members

. . . lauded by the American Psychiatric Association, which *represents* more than 36,000 physicians . . .

Employment history

Assigned to the Domestic Relations Court, later renamed Family Court, Bolin fought racial discrimination . . .

Kristy's project: Extracting numerical facts from news

- ✔ Identify and normalise numeric and date values
  - "...was forced to inject 183.5 million dollars into Aerolineas this year to keep it operating and pay its 9,000 employees."
  - "...was forced to inject 18350000 dollars into Aerolineas 2012 to keep it operating and pay its 9000 employees."
- Next step is to identify the type and slot each value fills

Slot	Type	Value
Investment	DOLLARS	183500000
Event time	YEAR	2012
Number of employees	NUMBER	9000

Capital Markets CRC Limited

Quote-based opinion mining

- Traditionally sentiment analysis has involved finding spans of text and labelling them as either positive or negative
- Sometimes this makes sense, e.g. positive about the carbon tax
- Other times it doesn't, e.g. positive about climate change
- What we really want are expressions of a point of view
- We aim to find quotes that represent different opinions

A more complex point of view

"However, the political feasibility of all countries agreeing to a harmonised carbon tax to achieve this outcome is highly questionable"

Capital Markets CRC Limited

Creating the data

- Getting agreement on what constitutes an opinion is very difficult
- Others have addressed this with lots of annotator training (approx. 40 hrs per annotator)
- We consider disagreement to be natural for this task
- All quotes have been double-annotated and should be triple-annotated by early next week
- Corpus statistics:
  - 700 documents covering 7 topics
  - 3141 quotes, average of 4.5 per doc
  - Estimated final cost for external annotators is $1200 (approx. $0.85 for an annotator to complete a doc)

Capital Markets CRC Limited

Modelling event references as links

- We can learn from some hyperlinks in Fairfax stories:
  - The body . . . was found under a plastic sheet . . .
  - . . . that the global hacking incident won't affect them.
  - . . . 29-year-old Andy Marshall died this week after . . .
  - Ellison was last year named best-paid executive
- and less so from others:
  - . . . the Mitsubishi i-MiEV, which is now on sale. . .
  - Read our previous Brisbane's Best: CBD lunches for $10 | Chips | Bakery | Mexican | Pizza
  - . . . according to the QS World University Rankings.
- We can filter some of these out with basic rules:
  - hyperlink density, punctuation, . . .

Capital Markets CRC Limited

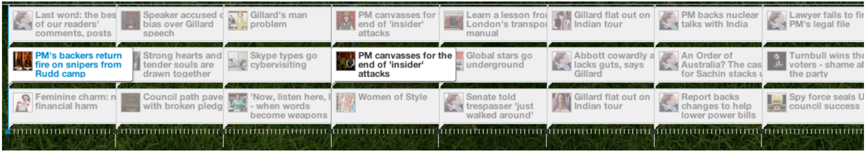
How can we predict hyperlink targets?

- Baseline: search archive with query = hyperlink anchor context
- Hypothesis: higher weight for target sentences with *new* content
- We index different portions of each article, such as:
  - Story text
  - Story title
  - First sentence
  - Sentences containing "yesterday", etc.
  - More sophisticated approaches...
- Learn from existing hyperlinks how to weight these "zones"
- So far this yields 17% relative MRR gain

Capital Markets CRC Limited

Concept timelines need to be selective

We can produce a list of documents mentioning Julia Gillard



But which are the most important to show on a timeline?

Capital Markets CRC Limited

Fergus's project: Choosing the best stories for a timeline

- Baseline approaches:
  - ✔ Cluster stories by text and timestamp, and choose a representative from each cluster
  - ✔ Borrow techniques from summarisation to select key stories
  - Use concept relevance scores to distinguish stories where Gillard is central or peripheral
- We would like to compare news stories to Wikipedia...

Capital Markets CRC Limited

Processing historical news

- Our extracted information enhances the future and the past.
- Four collections of Fairfax data:
  - DCDS (2008–present)
  - FutureTense (2002–2009)
  - NewsStore (1986–2009)
  - Microfiche via Google OCR (1830–1989)
- We would like to process these seamlessly.

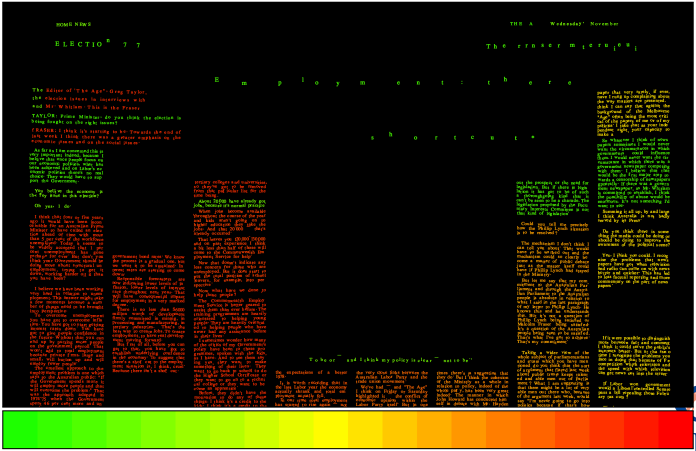
Towards a unified archive

- We can store the digitised texts in a unified format.
- Statistical models are best without duplicates:
  - We intend to implement automatic deduplication
  - Fairfax should provide any known duplicate IDs

Sebastian's project: Cleaning and ordering OCR output

- Segmenting images into logical segments (paragraphs, headings, bylines, etc.) to extract page features and reading order.
- This will enable us to extract individual articles from pages
- Google OCR data does not adequately group together contiguous segments of the page
- Our segmentation identifies paragraph and image segments reasonably well
- Next step: NLP on noisy text (Andrew's PhD)

Colour shows position in Google-generated HTML file



Capital Markets CRC Limited

Current segmentation progress

Supporting news timeline

- Store NEL data
- Check API calls
- Link and load 5-10 years or articles
- Link and load available photo and video assets

Improving linking performance with journalist feedback

F1	Method
n/a	Baseline: use only Wikipedia statistics
70.98	Train an NER model for SMH-specific expression of names
73.96	Use concept-mention cooccurrence statistics to bias toward commonly observed links from SMH articles

Further directions

- Concept-concept cooccurrence to bias toward concepts that occur together (requires `count2` data)
- Learn a statistical model for NEL

Batch annotation and correction

- **TODO: DT: we mentioned this last meeting, but nothing since**