

The Computable News Project
Milestone 8
NEL in the Real World

James Curran

ə-lab

School of Information Technologies
The University of Sydney



Computable News team

- Project leader: **James Curran**
- Project manager: **Candice Loxley**
- Postdoctoral researchers: **Daniel Tse**
- Senior Software Developer: **Will Cannings**
- PhD students:
 - **Joel Nothman (writing thesis)**
 - **Will Radford**
 - **Tim O'Keefe**
- PhD students on related work:
 - **Tim Dawborn**
 - **Andrew Naoum**
 - **Glen Pink**

Zoom has launched!



zoom

BETA

Zoom on Eddie Obeid

11:21AM Wednesday Apr 03, 2013 4,640 online now Do you know more about a story? Real Estate Cars Jobs Dating Newsletters Fairfax Media Network

MY NEWS MY CLIPPINGS MY COMMENTS MY HISTORY MY BENEFITS SIGN-UP LOG IN LEARN MORE

The Sydney Morning Herald


Where food & wine are made for each other

100% PURE NEW ZEALAND

News Sport Business Politics Comment Tech Entertainment Lifestyle Travel Cars Property Multimedia

Eddie Obeid
637 STORIES

What is Zoom?

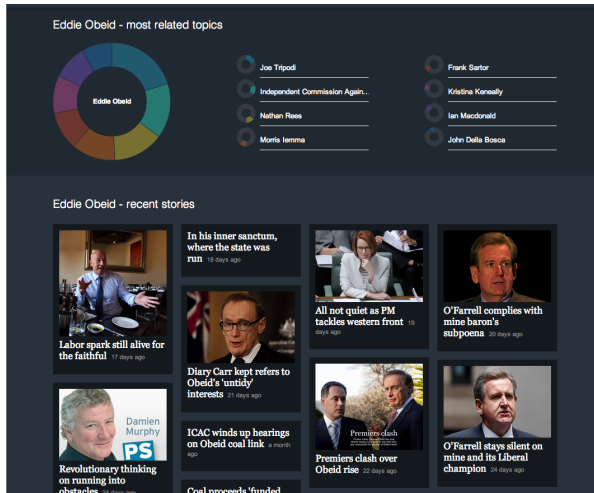


Nationals drop candidate
21 March 2013 | Richard Torbay has been dumped as the Nationals candidate for New England after information emerged that he had previously accepted funding from Labor to run as a state independent against Nationals candidates.
[Read this article](#)

See give riding fund-siser water...
Trigg takes rich behind murder to...
Ex-minister has record not many would...
Revolutionary thinking on running into...
O'Farrell stays silent on mine and...
O'Farrell complies with mine baron's...
All not quiet as PM backs western...
Labor spark still alive for the...
Nationals drop candidate

1990 1991 1992 1993 1994 1995 1996 1997 1998 1999 2000 2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013

Zoom on Eddie Obeid



Zoom on BHP Billiton

11:24AM Wednesday Apr 03, 2013 4,674 online now Do you know more about a story? Real Estate Cars Jobs Dating Newsletters Fairfax Media Network

MY NEWS MY CLIPPINGS MY COMMENTS MY HISTORY MY BENEFITS SIGN-UP LOG IN LEARN MORE

The Sydney Morning Herald

zoom BETA

The key to unlocking the secrets of a Marlborough Sauvignon Blanc is where you drink it

News Sport Business Politics Comment Tech Entertainment Lifestyle Travel Cars Property Multimedia

BHP Billiton 16,084 STORIES What is Zoom?

Goldman cuts iron-ore forecast as global demand growth slows

21 March 2013 | Goldman Sachs Group has cut its estimate for iron-ore prices this year on expectations demand will moderate and steel production will slow in China, the world's largest buyer.

Read this article

Soft season has signs of hope

Chinese miners ready to bulk up

Market yields to upturn

Chinese miners build muscle to compete...

Xairata chooses brown over green

Fake market signal? China slams big...

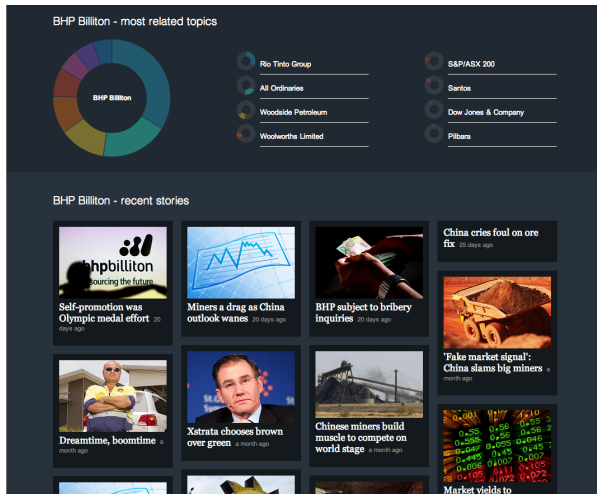
Stoke to watch on Wednesday

Mining's man with the Midas touch

Goldman cuts iron-ore forecast as global...

1995 1997 1999 2001 2003 2005 2007 2009 2011 2013

Zoom on BHP Billiton



M8: Engineering

- ✓✓✓ Supported Zoom development and launch
- ✓✓ Redesigned relevance metrics:
 - ✓✓ Concept-concept (replacing M1-7 metric)
 - ✓ Concept-story (replacing OpenCalais metric)
- ✓ Data cleaning and normalisation
 - ✓ Non-article filtering
 - ✓ Duplicate detection
- ✓✓ Improved linker precision
- ✓✓ Improved bulk linker speed
 - ✓ Created linking queue (for continuous updates)

M8: Research

- ✓✓ Two papers submitted to ACL (top international conference)
- ✓✓ Summer Scholarships complete
- ✓✓ Local description for NEL
 - ✓ Slot-filling for fact extraction
 - ✓ Quote-based opinion mining
 - ✓ Indirect speech corpus
 - ✓ OCRed historical archive (Google news data)

Datafest

- Helped support The Age Datafest hackathon
- Provided all available linked stories, counts and concept information through an API
- Great success with almost all teams using the Fizzing Panda data



Supported Zoom development and launch

- Collaboration++: onsite work, daily meetings, and fast responses
“Will Cannings --> Legend”
– George Wright
- Linking of the **entire archive** (back to 1986)
- Created a Cassandra/Python prototype API server (pre Rails)
- Helped with Cassandra, Rails and bits of Javascript/CSS
- Assisted with data loading and relinking
- Hosted the dev/testing server (Rails and latest link data)
- Created a linking queue to keep Zoom up to date

Zoom forced engagement with Fizzing Panda output

- Quantative evaluation (e.g. F-score) only tells part of the story
- Qualitative evaluation gives a more intuitive understanding
- Zoom has resulted in a deeper, qualitative, dive into the data
- realisation: $\sim 74\%$ F-score \Rightarrow 1 in 4 entities is wrong!
- Problems were discovered in the raw text, algorithms, and linker
 - raw data: duplicate (and near-duplicate) stories
 - raw data: non-stories (e.g. tenders, letters)
 - algorithms: irrelevant stories
 - linker: story “furniture” (e.g. Twitter links and wire attributions)
 - linker: over-zealous person name matching

Key Benefits of Fizzing Panda

- 1 Fizzing Panda is state-of-the-art: no-one else is much better
- 2 Fizzing Panda is not a black box
(cf. major relevance and linking changes during Zoom dev)
- 3 Tuning Fizzing Panda to local data is critical
- 4 This is as bad as Fizzing Panda will ever be because...
- 5 Fizzing Panda + journos is Fairfax's strategic advantage

Duplicate detection

- News Store, Future Tense and DCDS overlap
- Publication across up to 5 sites
- Stories evolve (and are tweaked)
- This is good until you want to count them
- We found duplicates within and across sources
 - different titles
 - different content
- Removed 387,250 duplicate stories out of ~3 million

Clustering by signature

- Create a signature for each document (in parallel)
 - title** normalized (lower-case, punctuation removed)
 - 5-gram** 20-hashed 5-grams normalized
- Cluster twice, once on each signature, reporting mean statistics over all story pairs (**s**, **t**) in the cluster
 - bow** cosine similarity between bag-of-words (normalized unigrams, top-20, no stopwords) for **s** and **t**
 - overlap** mean overlap of 5-grams for **s** and **t**

Filtering and postprocessing clusters

- Filter “true” clusters where
 - `title` bow ≥ 0.85 and overlap ≥ 0.9
 - `5-gram` overlap ≥ 0.95
- Identify the main story from each true cluster, preferring
 - earlier stories (using timestamp)
 - better sources (`dcds` \geq `ft` \geq `rly`)
 - larger story id (reissued)
- Hide non-main stories from the title and 5-gram clusters

Concept-concept relevance

- original metric always pulled up Australia, Sydney, NSW, ...
- new reverse relative frequency metric developed
- compared against raw freq, t-test, chi-squared, and PMI

Concept-story relevance

- Relevance is important from two perspectives
 - Determining which stories are most relevant to a concept
 - Determining which concepts are most relevant to a story
- Zoom needed a way to pick stories to appear in the timeline
- Articles need a way to pick concepts to link to
- First delivered an implementation of the OpenCalais relevance metric during milestone 5 – 22nd May 2012
- This metric tended to include incidental mentions, and resulted in unsuitable Zoom timelines

How do you determine relevance?

On a scale from one to five...

5085112

... but just when you thought it was safe to swim again

Kate Benson Medical Reporter

NINETEEN public swimming pools are being investigated by the Health Department after the number of people suffering a severe abdominal infection soared to more than 200 yesterday, the highest number recorded in February for a decade.

NSW Health fears Sydney is on the brink of an epidemic of cryptosporidiosis, a parasitic infection that causes cramping, diarrhoea, vomiting and fever, after another 44 cases were reported yesterday.

Half have been children.

The pools targeted included the Sydney Olympic Park Aquatic Centre at Homebush Bay, the Annette Kellerman Aquatic Centre in Enmore and the Manly Andrew Boy Charlton Swim Centre.

Three pools in the Blue Mountains, 11 in the western suburbs and two in southern NSW have also been named.

The director of communicable diseases at NSW Health, Jeremy McAnulty, said yesterday none of the 19 pools had tested positive for cryptosporidium but sufferers had reported swimming at those venues before becoming ill.

He said all the pools had been ordered urgently to "super chlorinate" using 10 milligrams of chlorine per litre in a bid to stem the rising rates of infection.

"The message here is that if you've got diarrhoea or had diarrhoea in the past two weeks, stay out of pools because you could contaminate that pool and cause many other cases of crypto, maybe even thousands," Dr McAnulty said.

The public swimming pool and spa pool guidelines, devised by NSW Health, advise that if a swimmer has a bout

Article

5085112

Save

Save and next

Back to list

☒ Document complete?

Quote Chains

3

Andrew Boy Charlton Swim Centre

3

Annette Kellerman Aquatic Centre

Ignore

Blue Mountains

Ignore

Enmore

Ignore

Health Department

Ignore

Homebush Bay

4

Jeremy McAnulty

Ignore

Kate Benson

Ignore

Manly

Ignore

NSW

1

NSW Health

Annotation

- Rank the concepts by relevance or mark as irrelevant
- Given the time constraints, the annotation scheme was essentially 'do what makes sense to you'
- A total of 3,222 annotations over 185 articles were produced in about 3 days
- Big thanks to Adam, George, Zoe and Candice for annotations
 - Adam powered through 570 annotations in one afternoon!

A statistical model of relevance

- We trained a linear regression model using the following features
 - TFIDF of the coreference chain
 - proportion of the document covered by the chain
- This evaluates at 0.082 MSE
- Other features were tested, but didn't perform as well as the two used in this model
- Better performance could be expected with a more rigid annotation scheme and more feature experimentation

Source and non-article filtering

- Through trial and error and lots of late night eyeballing we refined the set of data we link, and the way we treat the data
- Only stories from the SMH and The Age (and their Sunday papers) are included
- Stories with less than 2 paragraphs are ignored
- Reduced story count from 6.9M to ~3M

Non-article detection – excluded sections

- births, marriages, and deaths
- tenders
- the guide
- domain
- green guide
- supplement
- a2
- m
- letters to the editor
- letters
- the list
- news extra - letters to the editor
- news extra - letters
- obituaries
- column 8

Non-article detection – excluded titles

- today's flight schedules
- university of sydney examination awards
- suggestions, please
- regular shorts
- letters
- vice regal
- tv previews
- trivia
- short takes
- pay tv
- your letters
- correction
- corrections
- blocklines
- biblio file
- in brief and (in brief)

Linking changes

- More focus on precision rather than recall
 - Ignore 'noisy' alias sources
 - More restrictive handling of person names when searching
 - Better handling of location names
 - NER bug fixes
- Improved linking and data generation speed
 - Linking all 27 years of the archive previously took around 1 week
 - A full relink and data load can now be run in close to 7 hours
 - (on 4 dedicated 32 core machines. . .)

Linking changes

- Concepts mentioned less than 2 times in a story are ignored
 - Concepts mentioned in bylines are ignored
 - Some text shouldn't be linked
- **Follow WAtoday on Twitter**
- Newswire concepts are ignored if they appear in the last sentence

Milestone 9 TODOs

- NIL linking policy
 - Denis Naphine
 - Tom Waterhouse
- Abstract topic creation
 - Sydney shootings
 - Gay rights

Milestone 9: Search

- from linked entity set
 - canonical names
 - linked aliases
- full text search over articles using Solr
 - bounded time and section
 - integration with Zoom and known concepts

Milestone 9 TODOs

- understand and replace News Store
- get the correct home version of the article
- support Multi-tenancy
- keep Cassandra data up to date using linking queue

Part I

Computable News Research

Local description of entities

- Creation: journalists need to add context
- Disambiguation: readers need to identify the entity
- Curation: editors describe new entries in the entity store

People

per-role-left-np

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

per-role-appos

Robert Hughes, the Hey Dad! star, will plead not guilty ...

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

per-age-appos

John Smith, 82, ...

Organizations

org-acro

The Ministry of Defence (MOD) today announced ...

org-left-np

Brazilian miner giant **Companhia Vale do Rio Doce** ...

org-left-pos

Angola's **National Electoral Commission** ...

Locations

loc-right-comma-np

Toronto, New South Wales ...

loc-left-np

the small town of **Formosa** ...

loc-right-np

a village near **La Union** city in Zacapa province^a

^aThere are two bits of information here

Local description in the TAC 11 queries (n=2250)

KB (n=1250)	%	NIL (n=1250)	%
None	43	None	36
Info	57	Info	64
loc-comma-np	25	per-role-appos	16
per-role-np	10	per-role-np	14
org-acro	8	org-acro	8
per-role-appos	2	loc-comma	6
loc-np	2	org-np	4

What is Apposition?

- Syntax: adjacent NPs
- Semantics: often coreferred
- Pragmatics: often introduces new information

{John Ake}_h , {48}_a , {a former vice-president in charge of legal compliance at American Capital Management & Research Inc., in Houston,}_a , ...

Systems

- Fast, but dumb

Rule patterns of POS and NE tags

- Re-trained Berkeley Parser

LBP labelled syntactic parser

- Statistical models using syntactic and semantic features

Phrase ML over single phrases

Joint ML over pairs of phrases

Results submitted to ACL 13

Model	P	R	F
Rule	65.3	46.8	54.5
LBP	66.3	52.2	58.4
Phrase	74.6	44.4	55.7
Joint + LBP	69.6	51.6	59.3

Next

- Link using local description
- Investigate local descriptions for entity store population

Extracting facts for Slot Filling

- 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)
 - Relation/attribute types (slots)
 - Resolving duplicate or conflicting slot values
- Can be used to identify relations between entities in Zoom
- Currently replicating state-of-the-art

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.

Untyped relation extraction

Malcolm Turnbull has *ousted* Brendan Nelson as *leader of the federal Liberal Party*.

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

“ Nobody wants a new tax, but we have to do something about climate change”

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, so we triple-annotated all quotes
- Corpus statistics:
 - 700 documents covering 7 topics
 - 3,141 quotes, average of 4.5 per doc
 - Final cost for external annotators was \$1,300 (approx. \$0.93 for an annotator to complete a doc)
- A paper describing this corpus and its creation was submitted to ACL

Indirect speech

- We have also created a corpus of indirect speech covering 953 documents
- This will be used to train classifiers that can automatically extract all reported speech
- Corpus statistics:
 - 965 documents with more than 600,000 words
 - 7,973 quotes (average of 8.3 per doc)
 - 4,203 were direct, 2,921 indirect, and 849 mixed

Processing historical news

- **TODO: rewrite**
- Our extracted information enhances the future and the past.
- Four collections of Fairfax data:
 - DCDS (2008–present)
 - FutureTense (2003–2008)
 - RLAY (1986–2009)
 - Microfiche via Google OCR (1830–1989)
- We would like to process these seamlessly.

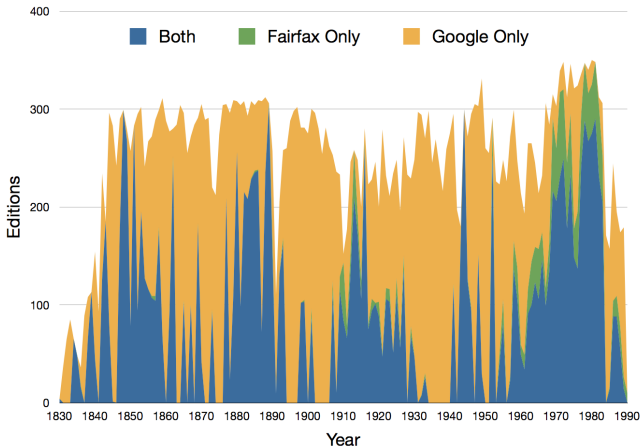
Towards a unified archive

- 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

Google OCR data

- We have cloned 17TB of data from Fairfax.
- Basic analysis:
 - 17,447 editions of The SMH from 1830 to 1990.
 - 6,369 editions of The Age from 1865 to 2001.
- In total, 362,646 scanned pages of text.
- Still a large amount of missing data: Google has 50,293 editions that we don't have.

Comparison of Fairfax and Google archives (SMH)



Sebastian's project: Cleaning and ordering OCR output

- Built a framework to represent the logical structure of a page accurately.
- Allows for quick and easy experimentation and substitution of various extraction algorithms.
- Implemented a line detection algorithm to improve page segmentation process.

Line detection and resulting segmentation

[illegible]