The Computable News Project
Milestone 2
Dynamic Entity Resource Pages

James Curran

ə-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**
- additional annotators from ə-lab and outside

Capital Markets CRC Limited

M2: Entity pages, event linking, web-sourced annotation

- Demo: dynamic entity pages with improved entities
- Access to new Fairfax Digital content feeds
- Event annotation schema design and testing
- Event annotation tool development
- Web-sourced entity annotation with Freelancer

Capital Markets CRC Limited

M2 Deliverables (high level)

- ✔✔ Demo: dynamic entity pages from SHM data
- ✔✔ 1,000 further SMH articles annotated with entities
  - ✔ Web service interface to named entity recogniser
  - ✔ Updated event annotation scheme design and testing
  - ✘ 500 further SMH articles annotated with events (250 complete)
  - ✔ Engine for generating a tag cloud of related entities
- ✔ ~~500 further SMH articles annotated with sentiment~~
- ✔✔ Web-sourced annotation using Freelancer

Capital Markets CRC Limited

M2 Demo

- ✔ Dynamic entity pages from Sydney Morning Herald data
  - iPad-friendly web interface as collection point for entity information
  - A further thousand SMH articles annotated with entities
  - Facts from Wikipedia (e.g., birth date, web site)
  - Entity mentions on Twitter & entity tweets
- Outsourced annotation from the web using Freelancer
  - Leverages powerful M1 annotation tool for web-sourcing
  - 3x increase in annotator cost efficiency
- Updated event task that links mentions to first stories
- Web service interface to named entity recogniser

Capital Markets CRC Limited

M2 Demo: Feature checklist

- ✔ a list of stories mentioning the entity, organized by:
  - ✔ date of publication
  - ✔ internal mentions of other entities and/or events
  - ✔ external mentions (e.g. on Wikipedia page hits, Twitter mentions)
  - ✔ related stories
- ✔ a list or tag cloud of related entities
- ✔ linking between dynamic entity pages
- ✔ links and/or content from other resources such as Wikipedia
- ✘ Twitter hashtags (used high quality usernames from Wikipedia)

Capital Markets CRC Limited

M2 Demo: Bonus features

- ✔✔ story images for entities based on caption text
- ✔✔ iPad-friendly demonstrator
- ✔✔ entity search (with auto suggestions)
- ✔✔ trending entities by type
- ✔✔ improved story view with entity browsing
- ✔✔ scaled demonstrator data 4x
  (from 2009 only to 2006-2009 plus recent)

Capital Markets CRC Limited

Data coverage

	1986/01-2010/01	2010/02-2011/04	2011/05-
Stories for all mastheads	✔	✘	✔
Comments, images, related assets	✘	✘	✔
Asset hit counts	✘	✘	✘

We are also collecting Twitter updates:

- mentioning smh.com.au (potential proxy for asset counts)
- by entities
- about popular entities

Capital Markets CRC Limited

What is a named entity?

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

What is a named entity?

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.

- Eddy Groves $\rightarrow$ wiki/Eddy_Groves
- ABC $\rightarrow$ wiki/ABC_Learning
- ABC $\nrightarrow$ wiki/Australian_Broadcasting_Corporation

Capital Markets CRC Limited

M2 Entities: deliverables

- ✔ Improved entity annotation tool
- ✔ 1,000 further SMH articles annotated with entities
- ✔ Engine for generating a list of stories about a given entity.
- ✔ Web service interface to named entity recogniser.

Capital Markets CRC Limited

M2 Entities: bonus

- ✔✔ Better type classification using Wikipedia pages
- ✔✔ Better recognition using SMH annotated data from M1
- ✔✔ More accurate linking
- ✔✔ Faster linking

Capital Markets CRC Limited

Annotation analysis

- Over 500k+500k words (over 320+254 man hours of work)
- 20,171 unique entities, 46% of them non-Wikipedia (M1+M2)
- 46,600 mentions, 29% of them non-Wikipedia (M1+M2)
- Type distribution:

44%	Individual
23%	Organisation
16%	Location
4%	Event
4%	Facility
4%	Work of Art
2%	Miscellaneous
2%	Product
1%	Artefact

Capital Markets CRC Limited

Event tracking

- ✔ Improved event annotation tool
- ✘ 500 further historical SMH articles annotated (250 done)
- ✔ Engine for generating a tag cloud of related entities

Capital Markets CRC Limited

What is an event?

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.

- Eddy Groves *is involved in a* JUSTICE-EVENT
- Eddy Groves *is a* CEO *of* ABC

Capital Markets CRC Limited

Existing event systems

- Aim to identify predetermined types of mentioned events, extracting who, when, where, etc.
- ACE 2005 corpus
  - Marks event "trigger" terms, participants, coreference
  - 34 types focussing on war, politics, business
  - Low annotator agreement (ambiguous); biased topic
  - Extracting all its information requires "semantic completeness"
- TimeML
  - Generic: most verbs and many nouns are marked as events
  - A more theoretical approach, interested in temporal relations
- OpenCalais (and others from late 1990s, early 2000s)
  - A small list of pre-configured event types
  - What, who, when identified with trigger words and patterns

Capital Markets CRC Limited

Event detection is hard to define and do

- Previous efforts have not achieved high annotation consistency or system accuracy, because:
  1. **Events don't have clear boundaries.** If a company lays off 100 workers, is that one EMPLOYMENT event, or 100? Is a house burning down in a bushfire part of the same event?
  2. **Events can be mentioned indirectly.** An article might quote a spokesperson saying We have increased profits by reducing our salary expenditures.
  3. **It's unclear what counts as an event.** Are price movements events? Does an event need a definite beginning and end? What about long-term events, like wars?

Capital Markets CRC Limited

Joel iterated through several event ideas

1. Like ACE, but coarser grained
   - Sentence level annotation
   - No participants
   - Coarser typology
2. Document level topic-based task
   - Maybe anchor words are the problem, so annotate at document level
   - Choose single event the article reports on
   - Usually can agree on event, type often a problem
   - Annotation too uninformative
3. Mark the article thread and annotate for type of new content:
   - Document level annotation
   - Assumes many articles belong to an identifiable 'chain' of stories
   - Assumes there's only so many types of stories

Capital Markets CRC Limited

Breakthrough: just link articles

- Defining and typologising events is hard
- Can usually agree when the paper refers to an event
  - ...But its properties or type are hard to define
- If paper references an event that may have been reported on, find the article and link it
- The boundary cases of maybe-an-event are very seldom specific or newsworthy
- But can we predict what will have articles?

Capital Markets CRC Limited

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

How Milestone 1 was annotated

- 111.75 hours by post-docs
- 54.25 hours by press-ganged PhD students
- 51.25 by project PhD students
- 51.25 friend hired at tutor rates

- Need to find a cheaper way...

Capital Markets CRC Limited

Amazon Mechanical Turk?

- Lots of interest in Amazon Mechanical Turk recently
- Marketplace for fine-grained labour tasks
- Turkers are assumed to be inter-changeable
- Little opportunity for detailed training
- Slightly mixed results, but most projects report success

Our project's a poor fit for Mechanical Turk

- It's best to annotate a whole document at once. Turkers prefer small tasks they can drift in and out of, so our granularity is a problem
- Annotators get more accurate on our task as they go, and may require feedback and corrections. Relationship with turkers assumed to be very minimal. Cannot contract a turker for some period of work.
- Our tool is complex and may be difficult to embed into MT. Tool critical for annotation.
- If we spend two weeks preparing, vetting and curating the data, was the annotation really very cheap?

Capital Markets CRC Limited

A better way: freelancer

- Traditional outsourcing lets you select, train and build a relationship with specific annotators
- Better incentives for cooperation: MT incentives often perverse
- Labour market still very competitive, so rates are affordable
- freelancer.com has very liquid market, excellent site and systems
- elance.com similar, but less liquid market, slightly worse set up

Capital Markets CRC Limited

Managing annotation projects

- Initial 'trial' projects of 10 to 15 hours of work
  - Stressed English language skills at top of job post
  - Advertised more work immediately to good annotators
  - Let them know we were looking to build relationships
- Good annotators were then offered further work

Capital Markets CRC Limited

Very competitive bids from USA

- For 250 docs, difference between $3 and $0.50 rates is $625
- Saving easily eroded by any extra management or QC cost
- USA bids were very strong
  - Accepted bids for $3 per doc, $2.25 per doc, $1.25 per doc
  - Responders advertised relevant experience, e.g. indexing academic books, English teacher, BA with linguistics credits, experience as a paralegal, etc
  - Articulate, well-thought out responses
- Developing world bids were less competitive
  - Plenty of $0.50 per document bids, but very few appealing
  - Bids with any language error automatically rejected
  - Most bids seemed cut-and-paste

Capital Markets CRC Limited

General Strategy

- We're betting reputation is good long-term currency
  - Most non-programming tasks are posted by evildoers (spammers, plagiarists, people cheating on their homework, etc)
  - Savvy workers will be good at tracking your reputation.
  - Reputation is cheaply lost and hard to regain
  - Don't skimp on rates, prompt payment, consistency, honesty...
- Training is costly if annotators leave
  - 3.5 hours training can be more expensive than a whole project
  - Must make sure annotators want to sell you all their time
- Best strategy: be nice to everyone!
- 16/16 100% positive employee reviews received

Capital Markets CRC Limited

Annotation costs

Employee type	Price per hour	Hours spent	Net price
Management (post-doc)	$50	80	$4000
Post-docs	$50	112	$5600
'Volunteer' PhD students	$35	54	$1890
Project PhD students	$40	51	$2040
Research assistant	$35	51	$1785
Milestone 1 total	$37	348	**$15315**
Management (post-doc)	$50	80	$4000
Freelancers	$10	250	$2500
Milestone 2 total	$18	330	**$6500**

Capital Markets CRC Limited