The Computable News Project
Milestone 9
???

James Curran

ə-lab
School of Information Technologies
The University of Sydney

26th June 2013

Capital Markets CRC Limited

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**
  - **Kellie Webster**

Capital Markets CRC Limited

M9: Engineering

- ✔ Zoom warranty
- ✔ ~~Living with Cassandra~~
- ✔ Hypertable and MapR benchmarking
- ✔ s3 concept and document storage
- ✔ Interim backend: Elasticsearch; MySQL; Hypertable ETL → s3

Capital Markets CRC Limited

M9: Learning from our mistakes

- ✔ Correction tools
  - ✔ Triage interface
  - ✔ "Purple" interface
- ✔ Applying fixes to `Zoom` stories

Capital Markets CRC Limited

M9: Products

- ✔ INSIGHT– Fizzing Panda Analytics
- ✔ `amp`– Games for the curious
- ✔ `shorty`– Summarising news
- ✔ `Zoom`
  - + videos
  - + sentiment
  - + groups
- ✔ `smhdiffs`– Tracking news
- ✔ Demoed at HackHackers

Capital Markets CRC Limited

M9: Research

- ✔✔ Two ~~long~~short papers ~~submitted~~accepted to ACL (top international conference)
  - ✔ Slot-filling / TAC
  - ✔ Local-description
  - ✔ Indirect speech corpus
  - ✔ OCR
  - ✔ Coreference

Capital Markets CRC Limited

Cassandra woes

- Discovered missing `count` and `concept` data
  - -2% accuracy
  - many missing concepts in `Zoom` data
- Annotation data went missing – 42% of writes failed
- Teleconference with DataStax suggested that `Cassandra` is inappropriate for our usecase.

Capital Markets CRC Limited

Cassandra's successor

- Benchmarked preferred database (`Hypertable`) reading and writing real world data
- Involved loading 1.7b rows of counts2 data – 40gb – and reading data back out
- On 2 servers, with 4 clients:
  - loading took 44m – 665k rows/s – with a peak speed of 930k/s
  - fastest sequential read speed: 5.5m rows/s
  - fastest random read speed: 2.2m rows/s
- Test was repeated by Fairfax on EC2 using `MySQL`
  - loading (inserts only, no counting) took around 4hrs
  - aggregation (counting step) hadn't finished after 72hrs

Capital Markets CRC Limited

Production backend plan

- `Elasticsearch` for full-text search
- `MySQL` for re-linking queue
- `s3` for rendered views
- `Hypertable` for concept, counts and other "big data"

Capital Markets CRC Limited

Product engineering

- Collection and linking INSIGHT sources
- Linking video OCR manifests
- Implemented baseline sentiment
- Implemented baseline summarisation
- Implemented groups
- Extensions to Lucene indexing allows smarter search over keywords and concepts

Capital Markets CRC Limited

Editors are part of the linking system

- Created a suite of web apps for Linker Quality Assurance

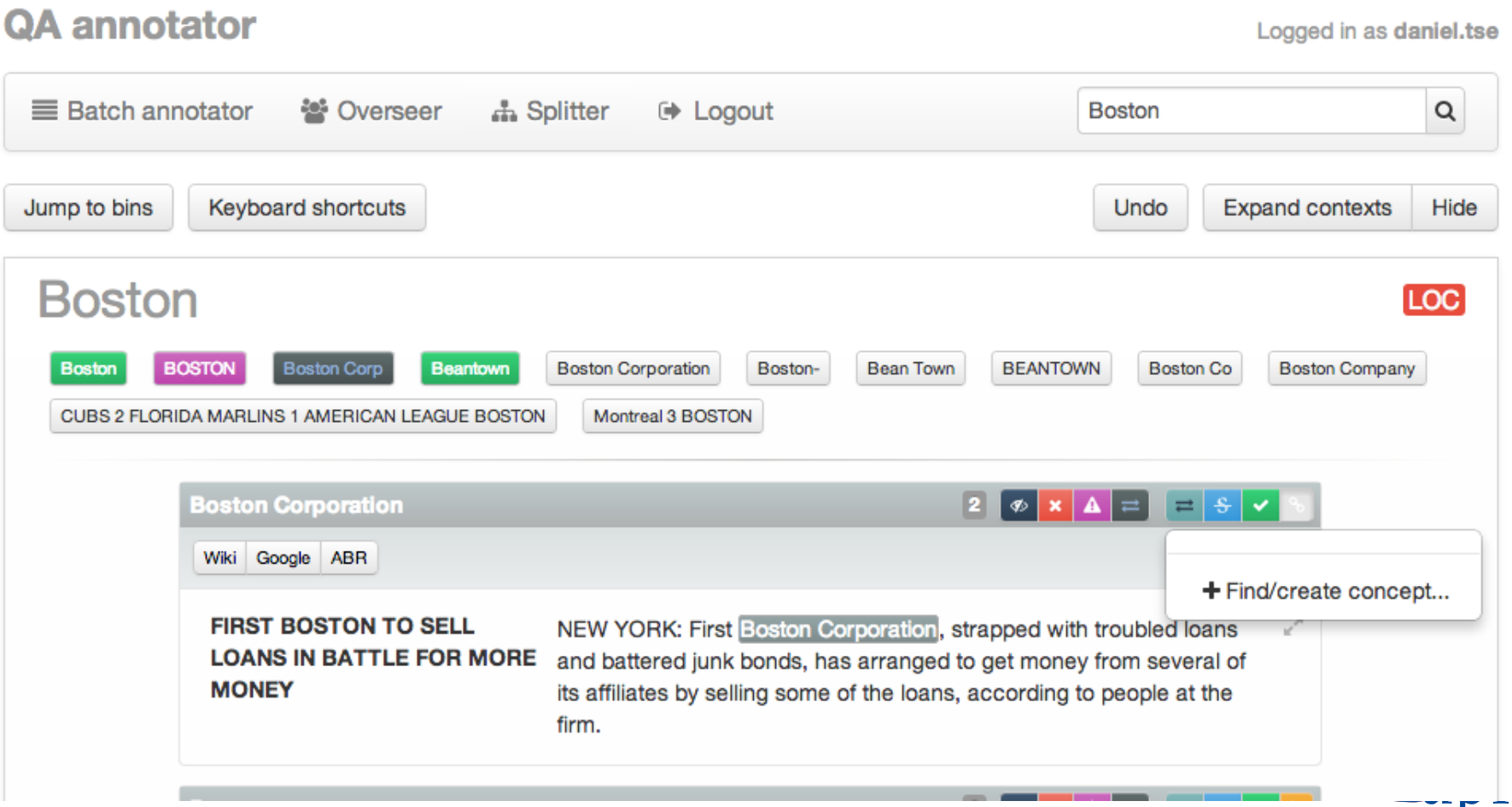
  Triage Centre lets editors pinpoint linking decisions that are broadly incorrect

  Concept Split is an advanced interface for fixing linking decisions at the document level

- The CompNews QA suite developed in this milestone:
  - allow rapid triage of linking errors and data collection
  - generate valuable data for improving the linker and NLP resources

Capital Markets CRC Limited

Triage Centre



Markets CRC Limited

Triage Centre: resolving linking errors

Don't link this to any concept	Mention text is garbage
Needs attention	Sends the case to the Concept Split interface
Mention boundary wrong	The boundaries of the mention need fixing
Spelling error	The linking decision is correct, but there's a typo
Correct	The linker did a good job!
Reassign to concept	The text should be linked to another concept

Capital Markets CRC Limited

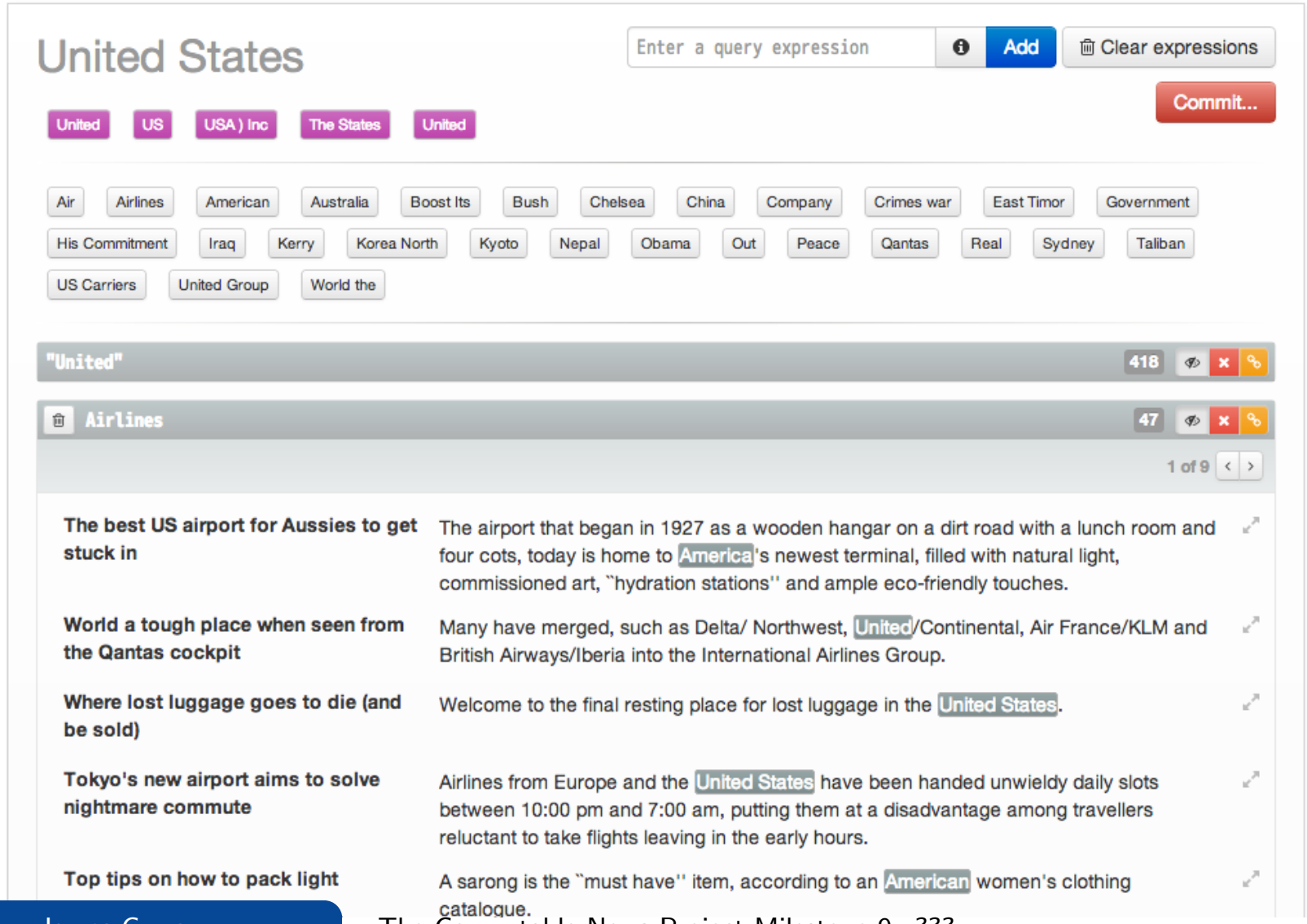
Concept Split

- Sometimes, a concept is mixed.
- `United` is sometimes correctly linked to United States, but sometimes should be linked to United Airlines.

- We created the **Concept Split** interface to split these difficult cases.
- Documents in which `United` should be linked to United Airlines are likely to mention air travel.
  - Find the subset of documents which mention `airfares`, and affect those documents only.

Capital Markets CRC Limited

Concept Split



Capital Markets CRC Limited

The Computable News Platform

- A platform allowing Fairfax to engage with readers with cutting-edge technology
- A small agile team able to deliver research quickly
- `Zoom` is just one product built on the platform

Capital Markets CRC Limited

INSIGHT- news analytics

- Live tweets: 60m+ total, 55 tweets/s live stream, 65 tweets/s SMH tweets
- Less than 15m article and comments delay: > 150k+ total

- `http://insight.schwa.org`

Capital Markets CRC Limited

amp- an experimental news game

- Pick a person and the 3 most-related people across time
- Guess who!
- `http://amp.schwa.org`

Capital Markets CRC Limited

shorty- entity-driven story summarisation

- Can we identify the most relevant content in stories?
- http://shorty.schwa.org

Capital Markets CRC Limited

Zoom extras

- Video from the QUT team
- Baseline concept sentiment
- Manually curated concept groups
  - Labor Politicians
  - NBN

Capital Markets CRC Limited

smhdiffs- how are stories changed?

- Various ways:
  - typos
  - updates
  - fact changes
- http://dev.willcannings.com/smh_diffs/

Capital Markets CRC Limited

Apposition

- ✔ Accepted into ACL
- Appositions are adjacent noun phrases that refer to the same entity:
  {John Ake} , {48} , {a former vice-president in charge of legal compliance at American Capital Management & Research Inc., in Houston}
- Extracting apposition could be very useful for slot filling and entity disambiguation
- We developed a new method that gets a >10% F-score improvement over previous work

Capital Markets CRC Limited

Opinions

- ✔ Accepted into ACL
- Defining opinions is challenging:
  - "Whether it's an ETS, whether it's a carbon tax, there will not be any new taxes as part of the Coalition's policy"
  - "I'm happy to see a debate about the nuclear option."
- Are these positive or negative? What do positive and negative mean here?
- It's easier to decide if a given quote agrees or disagrees with something
- So we provided a concrete expression to compare to: "Australia should introduce a tax on carbon or an emissions trading scheme to combat global warming."
- We annotated 7 topics, with 100 SMH articles per topic

Capital Markets CRC Limited

Opinion Annotation Tool

Quote Sentiment Annotation Annotation tasks Annotation guide

Business world fights coal levy

A levy on carbon fuels such as coal would be a disincentive to foreign investors , the executive director of the Business Council of Australia , Mr Paul Barratt , warned Federal Government ministers yesterday .

> `` The very fact that this matter is under discussion at all establishes an air of uncertainty about Australia as an investment location for energy intensive industry , "
>
> — Mr Paul Barratt

he said .

With regards to the topic statement "Australia should introduce a tax on carbon or an emissions trading scheme to combat global warming", could you use the quote to convince someone of the speaker's position

assuming the person is not knowledgable about the topic?

○ Strongly or clearly opposing ◉ Opposing ○ Neither supporting nor opposing ○ Supporting ○ Strongly or clearly supporting

assuming the person is knowledgable about the topic?

○ Strongly or clearly opposing ◉ Opposing ○ Neither supporting nor opposing ○ Supporting ○ Strongly or clearly supporting

I can't assign a sentiment to this quote because

☐ It is not a quote ☐ There is no sensible choice for the sentiment ☐ Is not related to the topic

I can choose a sentiment but this quote has

☐ an incorrect speaker ☐ an incorrect quote span

tal
ets
ited

Slot Filling/TAC

- Slot Filling shared task coming up in July
- Given an entity, find all values for particular attribute types
- Attribute types include **employee or member of**, **shareholders**
- Currently working on system for participation in task
- Approach is based on finding all possible potential values for entity, so that we can discover unknown attribute types (and avoid training the system specifically for each)

Capital Markets CRC Limited

Slot Filling examples

number of employees or members

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

employee or member of

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

Capital Markets CRC Limited

Local Description Update

- Last milestone introduced an annotation scheme over TAC data
- Some early linking results on CN stories, but requires supervised statistical model to overcome noise
- This is targetted at EMNLP'13 or PhD thesis

Capital Markets CRC Limited

Indirect Speech

- Quotes can be:
  - Direct: "The tax is a positive development," he said.
  - Mixed: The tax is a "positive development," he said.
  - Indirect: The tax is a positive development, he said.
- Previously we looked at extracting direct quotes and finding who said them
- This time we have annotated all direct, mixed, and indirect quotes (and their speakers) in 965 SMH articles
- We can extract the quotes with an f-score of 80%, and if we count partially correct matches we get 90%
- This will allow us to find most opinionated content within SMH articles
- We will submit this work to EMNLP'13

Capital Markets CRC Limited

Matching newspaper OCR with digital archives

- Matched 5,000 OCR pages with their digital counterparts.
- Vital training data for building an automated system to process the remaining 360,000 pages.

OCR text versus digital archives

Newspaper OCR

A season for cricket books . . .

WITH the Rugby League season upon us already, cricket books still lead the field in the bookshops, as this week's Top Ten shows.

But then, a good read about cricket is something you can enjoy at any season of the year.

A recent release which is certain to find favour is Jack McHarg's book **Stan McCabe** (Collins $14.95), on one of the greats of the Australian game.

Keen followers of the game often miss their reading if their eyesight gets weaker.

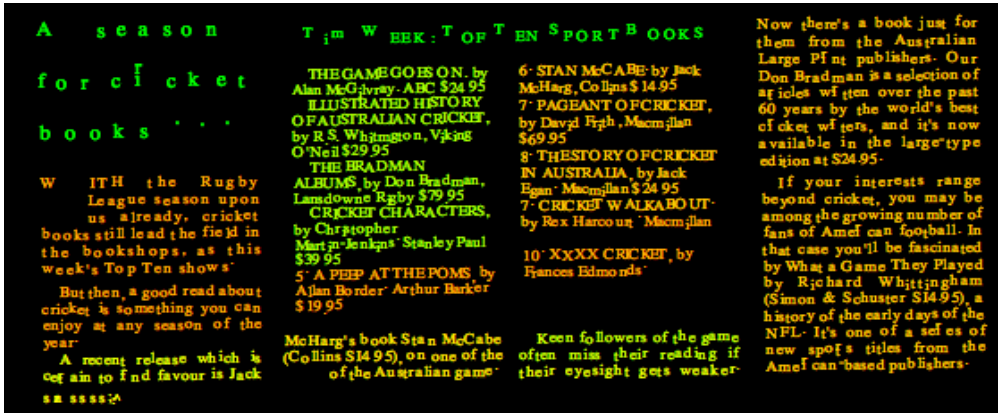
Now there's a book just for them from the Australian Large Print publishers. **Our Don Bradman** is a selection of articles written over the past 60 years by the world's best cricket writers, and it's now available in the large-type edition at $24.95.

If your interests range beyond cricket, you may be among the growing number of fans of American football. In that case you'll be fascinated by **What a Game They Played** by Richard Whittingham (Simon & Schuster $14.95), a history of the early days of the NFL. It's one of a series of new sports titles from the American-based publishers.

THIS WEEK: TOP TEN SPORT BOOKS

1. THE GAME GOES ON, by Alan McGilvray. ABC $24.95
2. ILLUSTRATED HISTORY OF AUSTRALIAN CRICKET, by R.S. Whitington, Viking O'Neil $29.95
3. THE BRADMAN ALBUMS, by Don Bradman, Lansdowne Rigby $79.95
4. CRICKET CHARACTERS, by Christopher Martin-Jenkins, Stanley Paul $39.95
5. A PEEP AT THE POMS, by Allan Border, Arthur Barker $19.95
6. STAN McCABE, by Jack McHarg, Collins $14.95
7. PAGEANT OF CRICKET, by David Frith, Macmillan $69.95
8. THE STORY OF CRICKET IN AUSTRALIA, by Jack Egan, Macmillan $24.95
9. CRICKET WALKABOUT, by Rex Harcourt, Macmillan $19.95
10. XXXX CRICKET, by Frances Edmonds, Heinemann $24.95

Compiled by Symons Bookshop



A season for cricket books . . .

WITH the Rugby League season upon us already, cricket books still lead the field in the bookshops, as this week's Top Ten shows.

But then, a good read about cricket is something you can enjoy at any season of the year.

A recent release which is certain to find favour is Jack

THIS WEEK: TOP TEN SPORT BOOKS

THE GAME GOES ON, by Alan McGilvray. ABC $24.95
ILLUSTRATED HISTORY OF AUSTRALIAN CRICKET, by R.S. Whitington, Viking O'Neil $29.95
THE BRADMAN ALBUMS, by Don Bradman, Lansdowne Rigby $79.95
CRICKET CHARACTERS, by Christopher Martin-Jenkins, Stanley Paul $39.95
5. A PEEP AT THE POMS, by Allan Border, Arthur Barker $19.95
6. STAN McCABE, by Jack McHarg, Collins $14.95
7. PAGEANT OF CRICKET, by David Frith, Macmillan $69.95
8. THE STORY OF CRICKET IN AUSTRALIA, by Jack Egan, Macmillan $24.95
9. CRICKET WALKABOUT, by Rex Harcourt, Macmillan
10. XXXX CRICKET, by Frances Edmonds

McHarg's book Stan McCabe (Collins $14.95), on one of the greats of the Australian game.

Keen followers of the game often miss their reading if their eyesight gets weaker.

Now there's a book just for them from the Australian Large Print publishers. Our Don Bradman is a selection of articles written over the past 60 years by the world's best cricket writers, and it's now available in the large-type edition at $24.95.

If your interests range beyond cricket, you may be among the growing number of fans of American football. In that case you'll be fascinated by What a Game They Played by Richard Whittingham (Simon & Schuster $14.95), a history of the early days of the NFL. It's one of a series of new sports titles from the American-based publishers.

Digital Archive

A SEASON FOR CRICKET BOOKS...

WITH the Rugby League season upon us already, cricket books still lead the field in the bookshops, as this week's Top Ten shows.

But then, a good read about cricket is something you can enjoy at any season of the year.

A recent release which is certain to find favour is Jack McHarg's book Stan McCabe (Collins $14.95), on one of the greats of the Australian game.

Keen followers of the game often miss their reading if their eyesight gets weaker. Now there's a book just for them from the Australian Large Print publishers. Our Don Bradman is a selection of articles written over the past 60 years by the world's best cricket writers, and it's now available in the large-type edition at $24.95.

If your interests range beyond cricket, you may be among the growing number of fans of American football. In that case you'll be fascinated by What a Game They Played by Richard Whittingham (Simon & Schuster $14.95), a history of the early days of the NFL. It's one of a series of new sports titles from the American-based publishers.

tal
ets
ited

Coreference

- cluster all references to an entity, including pronouns ('*she*') and common noun descriptions ('*the prime minister*')
- enrich information extraction
  - 'She will be defeated in September' said Opposition Leader Tony Abbott'
- currently buliding resolution system, based on re-implementation of Stanford's state of the art approach

Capital Markets CRC Limited