The Computable News Project
Milestone 10
Hypertable, Corrections and Multi-tenancy

James Curran

ə-lab
School of Information Technologies
The University of Sydney

21st October, 2013

Capital Markets CRC Limited

Computable News team

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

Capital Markets CRC Limited

M10: Summary of tasks

- ✔✔ Replacement of Cassandra with Hypertable
- ✔ Automatic application of corrections
- ✔✔ Multitenancy with NZ and Regional data
- ✔ Documentation

- ✔ Entity store editor
- ✔ Wikipedia concept import

Capital Markets CRC Limited

M10: Engineering

- ✔✔ Transition from `Cassandra` to `Hypertable`
  - ✔ So far, much faster and no dropped records!
- ✔✔ Substantial increase in linking speed

Capital Markets CRC Limited

M10: Multitenancy

- ✔ Concepts have tags for flexible multitenancy
- ✔ Documents have tags to produce multiple `Zooms`
- ✔ Linker uses tags to control available concepts
- ✔ Tag-specific correction rules

Capital Markets CRC Limited

M10: Correction tools

- ✔ Further work on correction tools
- ✔ Thousands of corrections entered
- ✔ Assess most frequent concepts and unlinked terms
- ✔ Supports multitenancy
- ✔ Fixes apply to `Zoom` stories

Capital Markets CRC Limited

M10: Concept editing

- ✔ Create and edit Entity Store concepts
- ✔ Import individual concepts from Wikipedia

Capital Markets CRC Limited

M10: Ongoing research

- ✔✔ Paper accepted to EMNLP (a top international conference)
- ✔ Paper submitted to ALTA (Australian workshop)
- ✔ Improved linker for TAC evaluation: still among the best
- ✔ Entered slot-filling system for TAC evaluation
- ✔ Submitted thesis on Event Linking

Capital Markets CRC Limited

M10: Wrapping up

- ✔ Compiling documentation in Google Docs
  - ✔ on the system
  - ✔ on project
- ✔ Will update instructions during installation (this week)
- ✔ `http://goo.gl/Bl5Nna`

Capital Markets CRC Limited

Farewell Cassandra, hello Hypertable

- Had found Cassandra unstable and lost many writes
- Found Hypertable fast and reliable
- Moved concepts, documents and statistics to Hypertable
- In-house ORM minimised changes to application code
- Refactored schemas to clean up accumulated problems
- Linking data compiled into efficient MsgPack blobs

Capital Markets CRC Limited

A challenging multitenant environment

- We want some data shared, and others site-specific
- Different products need to work with different sets of documents
- Different terms are prominent in different publications
- Some concepts may be special to specific sites
- Language may have different meanings according to site

Capital Markets CRC Limited

Flexible multitenancy through tags

- Each concept can be labelled with a set of tags: `all`, `wiki`, `regional`, `nz`, `fin`
- Linking requests specify the tags to link against
- Tags on documents define routing to outputs (e.g. `Zoom`)
- Corrections can be tag-specific

Capital Markets CRC Limited

Multitenant tags in correction

- Annotations are scoped to a tenant using tags
- Allows different tenants to provide different annotations
- The same term can be linked differently according to tenant

Capital Markets CRC Limited

Importing and editing concepts

- Import concepts from Wikipedia into store
- Create new concepts from scratch
- Edit existing concepts, e.g. to modify tags

Capital Markets CRC Limited

Viewing top terms

- List of most frequent concepts linked
  $\Rightarrow$ may need corrections due to high risk of viewing
- List of most frequent unlinked mentions
  $\Rightarrow$ may need new concepts to be created
- Can be scoped by date of publication, document category, etc.

Capital Markets CRC Limited

High-performing results at TAC evaluation

- Submitted systems with:
  - new work in local description matching
  - machine learning for candidate scoring
  - better NIL clustering by local attributes and topics
- 26 teams, 111 runs
- Top-class score, 1.6% below best system

Capital Markets CRC Limited

TAC results

System	All	KB	NIL
Top	72.1	72.4	72.0
1 Supervised with basic clustering	70.5	72.1	68.5
2 Supervised with local clustering	70.3	72.1	68.2
3 Supervised with topic clustering	68.9	72.1	64.8
4 Supervised +NIL, basic clustering	54.4	44.5	64.1
5 Supervised +NIL, topic clustering	53.0	44.5	60.9
Median	58.4	55.8	60.3

Capital Markets CRC Limited

TAC detail by genre and entity type

System	News	Web	Fora	PER	ORG	GPE
Top	80.1	67.3	63.3	75.8	73.7	72.0
1	77.0	63.3	63.3	73.4	67.6	70.3
Median	65.5	54.6	45.8	62.0	59.9	52.6

Capital Markets CRC Limited

Publication on quotation extraction

- Extracting and attributing direct, indirect and mixed quotations

Example

Police would only apply for the restrictions when "we have a lot of evidence that late night noise. . . is disturbing the residents of that neighbourhood", Superintendent Tony Cooke said.

- New data and techniques for extracting
- Published at Empirical Methods in NLP 2013

Capital Markets CRC Limited

Concluded experiments in Event Linking

- Events proved to be a very challenging space
- Joel submitted thesis on a new model of event reference
- Hyperlinks in Fairfax Digital stories are often good indicators of event reference
- Improved system to predict event-centred links

Capital Markets CRC Limited

Documenting Fizzing Panda

- Collating in Google Docs
- System description
- APIs and maintenance
- Underlying technologies
- Other areas pursued in the project

Capital Markets CRC Limited

Things we would like to see in the future

- Feedback corrections to supervised linking system
- Integrate new insights from TAC including improved NIL handling
- Mastering the scanned historical archive
- Application to personalisation / advertising
- Application to editing

Capital Markets CRC Limited