The Computable News Project
Milestone 5
Deployment, Engineering and Feedback

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** (starting in approx. 1 month)
  - second position still being filled (2-3 year position)
- Senior Software Developer: **Will Cannings**
- PhD students:
  - **Joel Nothman**
  - **Will Radford**
  - **Tim O'Keefe**
  - **Andrew Naoum** (starting in approx. 1 month)
  - **Glen Pink** (starting in approx. 1 month)
- PhD student on related work:
  - **Tim Dawborn**

Capital Markets CRC Limited

M5 Deliverables: Deployment and Engineering

- ✔✔ Deployed CompNews on Fairfax internal VM
- ✔✔ Delivered OpenCalais compatible API
  (planned for mid-milestone: 25th May, delivered: 22nd May)
  - ✔ Implemented document-scope concept relevance score
  - ✔ Implemented new topic linker (reducing software dependencies)
  - ✔ Analysed OpenCalais social tags and implemented first cut
- ✔✔ Completed Entity Store architecture
  - ✔ Benchmarked SQL/NoSQL solutions for counts data
  - ✔ Calculated data requirements (space and growth rate)
  - ✔ Confirmed Cassandra will meet requirements
  - ✔ Completed initial Cassandra data representation

Capital Markets CRC Limited

M5 Deliverables: Feedback API, research and bonuses

- ✔✔ Implemented user feedback API
  - ✔ Implemented live linking on demo (and dashboard)
- ✔✔ Implemented demo user interface (ready for dog-fooding)
- ✔✔ Bonus: OpenCalais/CompNews comparison on gold standard
  - ✔ Automated analysis against gold standard
  - ✔ Side-by-side view of all three versions
  - ✔ Manual analysis of errors on all three standards
- ✔ Bonus: Linking as you type prototype
- ✔ Research deliverables:
  - ✔ Event linking task is now well defined
  - ✔ Entity linking paper (Will R, Joel) accepted into AI Journal
  - ✔ Event linking paper (Joel) accepted into ACL
  - ✔ Quote extraction paper (Tim) accepted into EMNLP

Capital Markets CRC Limited

CompNews deployment at Fairfax

- M4: Fully documented installation process and dependencies
- Goal is to substitute OpenCalais with CompNews seamlessly
- M5: get Fairfax developers programming against CompNews API
- Planned for mid-milestone: 25th May, delivered: 22nd May
- Delivered a version of j-calais library that uses our API
  drop in substitute for existing library

- Added missing OpenCalais features to CompNews
  - Document level relevance scores for concepts
  - Topic linker (reducing software dependencies)
  - Analysed social tags and implemented first attempt

  (required for MyMasthead/video recommenders)

Capital Markets CRC Limited

Relevance scores for concepts

- OpenCalais returns the relevance of each concept in a document
- OpenCalais approach is (deliberately?) poorly documented
- We have implemented the standard TF-IDF approach:
  - TF: term freq. (# times concept appears in document)
  - IDF: inverse document freq. (# documents concept appears in)
- Next step is application-specific feedback from Fairfax

Capital Markets CRC Limited

Topic linker

- M4: Topic linker used WikMiner (from Waikato University)
- WikiMiner slow and introduced lots of extra dependencies:
  - separately processed Wikipedia dump and raw data files
  - separate MySQL database
  - separate Apache Tomcat instance
- We have implemented a new efficient Trie-based topic linker
- Heuristically links all non-entity concepts for Wikipedia
- Next step is application-specific feedback from Fairfax

Capital Markets CRC Limited

Social tags: analysis and implementation

- Plain English labels that describe in what a document is about
- OpenCalais approach is (deliberately?) poorly documented
- Our analysis of OpenCalais output found a mixture of
  - Salient entities and topics (according to relevance scores)
  - Relevant Wikipedia categories (of corresponding concepts)
  - Top-level IPTC NewsML categorisations
- We have implemented social tags using:
  - Entities: we extract salient named entities with relevance scores
  - Topics: we extract explicitly mentioned non-named entity topics
- Next step is application-specific feedback from Fairfax, and:
  - Categorisation into broad topical domains (need annotated data)
  - Adding broader Wikipedia categories from salient concepts

For more information, see this wiki page.

Capital Markets CRC Limited

Entity Store architecture

- Entity Store is central to the whole CompNews project
- It stores everything we know about entities/topics (concepts):
  - everything we extract from Wikipedia and other KBs
  - everything we extract from Fairfax news stories, Hansard, ...
  - everything journalists add/correct about each concept
- It stores every connection between concepts and stories
  - every link between a story and a concept
  - every statistical relationship between two concepts
  - temporal trends for every concept (including Wikipedia hits)
- It will be used enterprise-wide across Fairfax Media
- Scalability and flexibility are therefore critical issues

Capital Markets CRC Limited

Entity Store requirements

- Support linguistic annotations on story and Wikipedia text
- Support arbitrary meta-data associated with each concept
- Support relationships between concepts and stories
- Support temporal trend data about concepts
- Support statistical relationships between concepts
- Support (fuzzy) information retrieval queries on concept names

Capital Markets CRC Limited

Calculating data requirements

- Use our gold-standard annotated data to estimate requirements
- For each document, the Entity Store will, on average, store:
  - 22 new concepts discovered in the story (Concepts table)
  - 40 links between the story and concept (Links table)
  - 160 concept counts for day, month, year and total (Counts1 table)
  - 8400 concept pair counts for day . . . total (Counts2 table)
- Counts2 is the big bottleneck per document
- Future versions of CompNews will exploit Counts2 for linking

Capital Markets CRC Limited

CompNews: designed for Australian news

- OpenCalais is a free API for NLP
- However, it is:
  - black-box
  - not-localised to Australian content
  - owned by Thompson-Reuters
- An in-house, customised NLP platform will perform more accurately than OpenCalais for MyMasthead
- **Outcome** CompNews performs better because it finds more of the entities in text (higher recall) even if it isn't always as accurate (lower precision)

Capital Markets CRC Limited

OpenCalais and CompNews: apples and oranges?

- OC only links some locations, companies and electronics – **we only evaluate over ner boundaries and types**
- OC also links `Position`, `EmailAddress`, `MarketIndex`, `MedicalTreatment`, `URL`, `PhoneNumber`, `MedicalCondition`, `Currency`, `IndustryTerm` – **these are filtered out**
- OC performs pronominal coreference – **we filter out pronouns**
- OC and our system inconsistently include "the" in entity spans – **initial "the" tokens are removed from all output**
- OC and our system inconsistently handle topic spans – **any linked span starting with a lower-case character is removed from all output**
- We believe that this is the fairest NER evaluation possible

Capital Markets CRC Limited

API results over the cn-2-dev gold-standard data

API	Metric	Precision	Recall	F-score
OpenCalais	Mention	88.7	60.2	71.7
OpenCalais	Entity	**81.2**	55.1	65.6
M5 trie topic linker	Mention	**89.9**	**88.3**	**89.1**
M5 trie topic linker	Entity	74.8	**73.4**	**74.1**

- Mention evaluation compares how many mention strings (without context) each API returns wrt. the gold standard
- Entity evaluation combines the mentions with the entity type

Capital Markets CRC Limited

Type-analysis over same dataset

API	Type	P	R	F
OpenCalais	PER	**90.6**	72.2	80.4
OpenCalais	ORG	**73.8**	44.1	55.2
OpenCalais	LOC	74.8	55.2	63.5
OpenCalais	MISC	**55.6**	13.9	22.3
OpenCalais	PER/ORG/LOC	**82.1**	59.5	69.0
M5 trie topic linker	PER	83.8	**83.7**	**83.8**
M5 trie topic linker	ORG	72.1	**73.1**	**72.6**
M5 trie topic linker	LOC	**76.2**	**71.9**	**74.0**
M5 trie topic linker	MISC	38.9	**35.9**	**37.3**
M5 trie topic linker	PER/ORG/LOC	78.4	**77.4**	**77.9**

Capital Markets CRC Limited

Errors: entities that systems missed and spuriously marked

OpenCalais

```
56 Spuri US
27 Miss NSW
20 Miss Government
15 Spuri GM
 9 Miss Labor
 8 Miss Prime Minister
 8 Miss Parliament
 8 Miss Opposition
 7 Miss Randwick
 6 Miss Sydney
```

Compnews

```
59 Spuri US
15 Spuri GM
10 Miss Prime Minister
 6 Spuri Generation
 6 Miss Government
 5 Spuri president
 5 Spuri EBITDA
 4 Spuri Zack
 4 Spuri NSW
 4 Spuri Nice
```

For more information, see this wiki page.

Capital Markets CRC Limited

Manual analysis of OpenCalais and CompNews

- We have manually analysed 30 example stories
- Counted errors in gold, OpenCalais and CompNews versions
- Errors are either **missing** or **wrong** (incorrect type or span)

Corpus	Missing	Wrong
Gold standard	18	106
OpenCalais output	291	76
CompNews output	34	165

- OpenCalais misses a large number of entities
- majority of CompNews errors are incorrect entity type

Capital Markets CRC Limited

Event Linking presented at ACL 2012

Sydney man carjacked at knifepoint
There has been another carjacking in Sydney, two weeks after two people were stabbed in their cars in separate incidents. A 32-year-old driver was walking to his station wagon about 4.30pm yesterday when a man armed with a knife grabbed him and told him to hand over his car keys and mobile phone, police said. The carjacker then drove the black 2008 Holden Commodore. . . He was described as a 175-centimetre-tall. . . Police warned Sydney drivers to keep their car doors locked after two stabbings this month. On September 4, a 40-year-old man was stabbed when three men tried to steal his car. The next day, a 25-year-old woman was stabbed. . .

- What events make up in the news?
- How do you decide if two references are the same event?
- We present our approach
  - task description
  - annotated corpus
  - task analysis

Capital Markets CRC Limited

Quotes Research

- Our quote attribution established a new state-of-the-art, which will be presented at Empirical Methods in Natural Language Processing 2012
- Previous work has treated quote attribution as a classification problem
  - Given a quote, find the most probable speaker
- We treat the task as a *sequence* labelling problem
  - Given a sequence of quotes, we need to find the most probable *sequence* of speakers for those quotes
- This allows our system to make trade-offs between the current decision and previous decisions

Capital Markets CRC Limited

Quote Example

Lacey says: "Some British newspapers hailed the result as a 'victory' for Elizabeth II, but she did not see it that way. ..."
The national director of the Australian Republican Movement, Terrie James, said: " My comment is that she has a better understanding of the issue than some of the monarchists. ... another referendum for five plus years."
Lacey told the Herald yesterday it was important to distinguish between what Prince Philip thought and said and what the Queen thought and said, though naturally he would have an influence on her.
"The Prince was implying that he thought there was no doubt that Australia should become a republic,"

Capital Markets CRC Limited

Text Analysis Conference - Knowledge Base Population

- International shared task concerned with i) Entity Linking and ii) Slot Filling
- Main competitive venue for Entity Linking and the 2012 evaluation is at the end of August
  - Companies: Microsoft, LCC, DBPedia Spotlight
  - Universities: Stanford, CUNY, JHU, UIC, NUS
- We've been successful in the past, but submitted a minimal system last year
  - 2010 placed 2nd: current linkers based on this system
  - 2011 placed 9th: minimal changes from 2010 system
- Good performance in 2012 means:
  - accuracy improvements we can use in our deployed systems
  - publicity

Capital Markets CRC Limited