The Computable News Project
Milestone 4
Web API and accuracy improvements

James Curran

ə-lab
School of Information Technologies
The University of Sydney

Computable News team

- postdoctoral fellows have left in the last 3-4 months:
  - Ben Hachey (Thomson Reuters R&D)
  - Matthew Honnibal (Macquarie University)
  - David Vadas (Susquehanna)
- Project manager:
  - **Tim Yeates**
- PhD students:
  - **Joel Nothman**
  - **Will Radford**
  - **Tim O'Keefe**
- PhD student on related work:
  - **Tim Dawborn**
- Web designer:
  - **Will Cannings**

Capital Markets CRC Limited

M5-M8 prospects and schedule

- will still take some time to hire qualified postdocs
- esp. to replace Ben as a 3-year postdoctoral fellow

- blocking on CMCRC/USyd and Fairfax/CMCRC contracts (and then lengthy University of Sydney hiring process)

- hiring casual developers in the meantime to maintain momentum
- project manager/admin will be hired ASAP
- 1 postdoc from nearly complete PhD students (Daniel Tse)

Capital Markets CRC Limited

M4 and M5 switch

- original M4 was *User-driven Types and Alerting*
- original M5 was *Web Application Programming Interface*

- context made switching milestones better option:
  - substantial effort spent selling M5-M8 within Fairfax
  - substantial (ongoing) effort on contracts and hiring
  - install working system in case project completed at M4
  - loss of postdocs during M3/M4 due to job uncertainty

  - user-driven alerting already in MyMasthead
    $\Rightarrow$ entity groups are (most?) important next step

Capital Markets CRC Limited

M4 Deliverables and Research

- ✔✔ cloud deployment: COMPNEWS system on an EC2 instance
  - ✘ working installation in Fairfax (stalled on syndication)
  - ✔ Application Programming Interface

- ✔ graph-based Named Entity Linking
- ✔ label Propagation for NEL
- ✔ results of TAC 11 NEL shared task

- ✔✔ 800 SMH articles quote annotated, 400 double annotated
  - ✔ quality control
  - ✔ two additional datasets
  - Machine learning approach

- ✔ Wikipedia count data

Amazon EC2 and Fairfax deployments

✔✔ Fully documented installation process and dependencies

✔ EC2 `m1.xlarge` instance used as a test case

- learnt the true extent(!) of the software/data dependencies
- learnt (some of) the complexities of running cloud instances
- $\sim \$550$/mth/instance (using ephemeral storage for DB/feeds)

✘ onsite installation with David Gillies
(complete except for HTTP authentication errors on feeds)

Capital Markets CRC Limited

Topics for MyMasthead

- we've supplied a list of "clean" topics for MyMasthead
- frequency analysis of top-ranked Wikipedia topic links

Application Programming Interface

- ✔ named entity linking on the fly:
  - `/extract` – to JSON
  - `/markup` – to HTML fragment
- ✔ search for entities in the entity store:
  - `/search` – to JSON
- ✔ existing HTML output for different components
  - `/concept` – one or more (intersection) concepts in HTML
  - `/timeline` – one or more (intersection) concepts in HTML
  - `/story` – story text rendered in HTML
  - `/image` – image resources from the SMH syndication
- ✘ groups of entities and some JSON interfaces
- ✘ OpenCalais output format

Capital Markets CRC Limited

NEL: Overview

- M3
  - Better in-document coreference with Title heuristics
  - Whole-document evaluation
  - Error analysis tools
  - Engineering
- M4
  - TAC 11
  - Graph-based NEL
  - Label Propagation

7 days of TAC 11 → #9 of 21 teams

- New task: clustering NIL entities
  - John Smith (a)
  - John Smith (b)

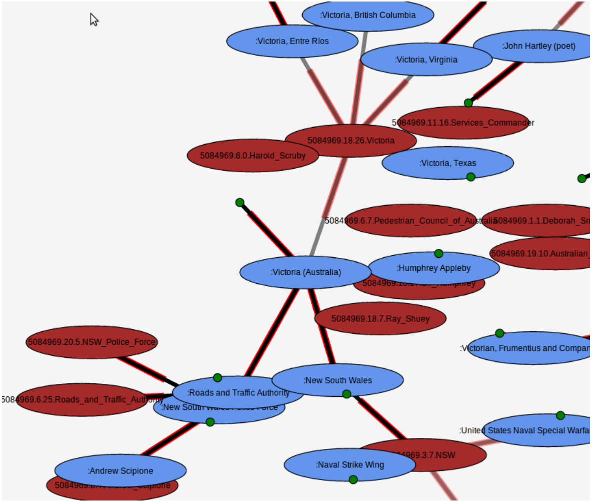
System	Accuracy	B^3 F-score
Monahan TAC 11 (LCC)	86.1	84.6
Cucerzan TAC 11 (MICROSOFT)	86.8	84.1
Zhang TAC 11 (NUS)	86.3	83.1
Cassidy TAC 11 (CUNY)	–	76.3
Chang TAC 11 (STANFORD)	79.0	76.3
Ratinov TAC 11 (ILL)	78.7	76.1
Anastacio TAC 11 (DMIR)	–	76.0
COMPNEWS TAC 11	**77.9**	**75.4**
Median	–	71.6

Capital Markets CRC Limited

TAC 11: What's new and what works

- Statistical classifiers for name variation (NUS) for 15% accuracy increase in acronym expansion.
- Topics: LDA (NUS), category and lexico-syntactic patterns (MICROSOFT).
- Query classification: different models for different entity types (DMIR).
- Cross-document: global consistency (MICROSOFT), clusters of documents (CUNY).

A graph of entities and mentions



- Vertices
  - Ment
  - Ent
- Edges
  - Ent → Ent
  - Ent → Ment

WISE 2011: Graph-based NEL- Hachey et al.

- Build a graph of entities and mentions from a document.
- Apply PageRank to the graph
  - A vertex's importance is based on the importance of its neighbours.
  - Random surfer interpretation.
- **"Topical" entities are ranked higher.**

WISE 2011: Competitive with the top systems

System	Web	Research	Accuracy
COMPNEWS TAC 10	✘	✔	84.4
PageRank	✘	✔	**85.5**
Lehmann TAC 10 (Unsupervised)	✘	✘	85.8
Lehmann TAC 10 (Supervised)	✔	✘	86.8
Zhang IJCNLP 11	✘	✔	87.6
Lehmann TAC 11 (Supervised)	✔	✘	89.8
Cucerzan TAC 11	?	✘	90.0

Table: Evaluation over TAC 10 test data

Label Propagation

- PageRank distributes one probability mass around a graph.
- A vertex holds a distribution of different labels based on those of its neighbours.
- Alternative sources of evidence for the linking decision.
- Parallelisable: each vertex calculation needs only its neighbours.

Case study: Recommending YouTube videos to users

- Partially labelled graph
  - Videos (labelled)
  - Users (unlabelled)
- Push labels from videos to users
- Users end with a distribution $\rightarrow$ ranked list of videos.

NER CoNLL 2003

System	Evaluation dev F	Evaluation test F	Training RAM	Training Time
C&C	88.98	83.91	0.3G	0h15m
new C&C	92.49	87.90	0.4G	0h20m
LBJ	93.50	90.50	14.0G	∼8h00m
Stanford	92.99	87.94	15.6G	∼4h00m

Capital Markets CRC Limited

NER M3 Data

System	Evaluation			Training		Tagging	
	P	R	F	RAM	Time	RAM	Time
C&C			73.71	0.90G	0h25m	0.33G	0m03s
new C&C	75.61	74.14	74.87	1.10G	0h25m	0.40G	0m03s
LBJ	78.83	78.21	78.52	9.70G	6h20m	3.00G	2m30s
Stanford	✘	✘	✘	✘	✘	✘	✘

NER M4 Data

System	Evaluation P	Evaluation R	Evaluation F	Training RAM	Training Time	Tagging RAM	Tagging Time
C&C	79.36	77.22	78.27	0.96G	0h30m	0.30G	0m04s
new C&C	80.15	78.55	79.34	1.20G	0h30m	0.50G	0m04s
LBJ	84.31	82.56	83.43	10.05G	9h21m	3.40G	2m54s
Stanford	✗	✗	✗	✗	✗	✗	✗

NER: Improvements

- New multi-word gazetteer features
- New larger Wikipedia-derived gazetteers
- Incorporated word-type information (Brown clusters)
- New contextual features
- Tweaking of training parameters
- Bug-fixes

Capital Markets CRC Limited

NER: Next

- Implement document-based features
  - Incorporate document structure information – are we looking at paragraph text, the article title, the byline?
  - Add support for acronym coreference
  - Add better support for all-caps sentences (these currently perform very poorly)
- Revisit tokenization

Quotes: M3 recap

- Quote annotation tool
- 800 documents annotated with quotes
- Quote extraction tool
- Rule-based quote attribution tool

Quotes: M4 goals

- ✔ Doubly annotate 400 of the already annotated 800 quote documents
- ✔ Determine quality of CompNews quotation corpus
- ✔ Test our software on other corpora
- ✘ Machine learning approach

Quotes: Motivating example

The opposition spokesman on climate action, Greg Hunt, said: ''This issue will be resolved well before 2016. If the Coalition is elected on the basis of scrapping the carbon tax, Labor must support its removal, including voting for its abolition in the House of Representatives and the Senate.''

Quotes: Cost-effective Freelancer annotation

- In Milestone 3, we collected 800 documents of quotation annotation for $700
- In Milestone 4, we continued to use Freelancer.com to double annotate 400 of those documents for $265
- All the annotators we employed for Milestone 4 had worked for us in previous milestones
- Most annotators were from the US

Quotes: Evaluation of corpus quality

- In Milestone 3, annotation quality was acheived through monitoring
- This tells us nothing about how hard the task is
- Double annotating documents allows us to calculate the agreement between annotators
- Average agreement over the 400 double annotated documents was **extremely high at 98.3%**
- Nobody has checked this for news text before

Capital Markets CRC Limited

Quotes: Two other corpora

- The first contains quotes from the Wall Street Journal
- The second contains quotes from late 19th century literature
- These will help us understand cases where our tools go wrong

Quotes: Rule-based approach

- The entity saying a quote is almost always introduced before the quote, or in the sentence where the quote ends
- Most quotes are also attributed using a reported speech verb ("said", "shouted", "exclaimed", etc)
- If there is a reported speech verb and an entity close to the quote, then we attribute the quote to the entity
- Otherwise we attribute the quote to the most recently mentioned entity
- For pronominal mentions we restrict the entity to be the gender that the pronoun implies

Capital Markets CRC Limited

Quotes: Results (gold standard)

- Quotation extraction is 99.34% accurate, with the few errors coming from missing quotation marks
- Quotation attribution performance over the three corpora is given below:

Corpus	Source	# Quotes	Accuracy
CompNews	SMH articles	3535	93.2%
Edinburgh	WSJ articles	3124	85.5%
Columbia	Literature	3064	56.7%

Capital Markets CRC Limited

Quotes: Machine learning approach

- We are recreating the results from a recent paper in the field that used machine learning
- Their work included some unrealistic assumptions, which we will be correcting
- Machine learning will let us take advantage of more features of the text including:
  - Who is mentioned inside the quote
  - How many recent quotes were made by an entity
  - Are there any other entities mentioned near the quote
  - Are there any other quotes near the quote
  - And many more...
- These features should let us improve the accuracy of out attribution system

Capital Markets CRC Limited

Quotes: Lessons learned so far

- Quote attribution accuracy is highly dependent on linker accuracy
- News text tends to be highly regular, making rule-based approaches quite accurate
- Machine learning improved the accuracy of other researcher's systems
- We're aiming to present this work at the 2012 ACL conference in collaboration with researchers in Edinburgh

Additional knowledge from Wikipedia

- terabytes of count data collected from 12/2007
- understand *reader* rather than writer popularity
- will be able to include hourly updates

- Wikipedia is now generating daily diffs
- we can now move beyond static snapshots of Wikipedia

Part I

Extra slides

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

Event Linking Scheme (1)

1. Find an event-denoting expression
2. Ignore it if:
    - hypothetical or uncertain
    - Not newsworthy, including:
    - Wrong semantic class (reporting, perception, etc.)
    - Non-newsworthy occurrence
3. Otherwise:
    - Select a single word expressing the event
    - If you have already marked another mention for the same event (or a closely related event first reported in the same article):
        - If that mention is in the same sentence, ignore it.
        - If that mention is in another sentence, mark the new mention as part of the same event.
    - Otherwise, mark the new mention as a new event.

Event Linking Scheme (2)

4. Select a category for the event:
   - Basic event – probably first reported in one news article
   - Complex event – likely to have multiple articles; often a named event
   - Trend or measured change
   - Many specific events
   - Non-specific
5. If a basic event:
   - Try to link to the article first reporting that event as having happened
   - Or mark as:
     - First reported here
     - Precedes 1986
     - Not found, which includes: No mention in archive, Not reported in archive, Not reported after occurrence
6. If a complex event:
   - Try to link to a Wikipedia article specifically about the event
   - ...k as Not found

Capital Markets CRC Limited

Annotation costs

Employee type	Price per hour	Hours spent	Net price
Post-docs	$50	112	$5,600
'Volunteer' PhD students	$35	54	$1,890
Project PhD students	$40	51	$2,040
Research assistant	$35	51	$1,785
Milestone 1 total	$42.2	348	**$11,315**
Milestone 2 total	$10	250	**$2,500**
Milestone 3 total	$7	100	**$700**