The Computable News Project
Milestone 6
Cassandra, TAC 2012 and Docrep

James Curran

ə-lab
School of Information Technologies
The University of Sydney

Computable News team

- Project leader: **James Curran**
- Project manager: **Candice Loxley**
- Postdoctoral researchers:
  - **Daniel Tse**
  - second position still being filled (2 year position)
- Senior Software Developer: **Will Cannings**
- PhD students:
  - **Joel Nothman**
  - **Will Radford**
  - **Tim O'Keefe**
- PhD students on related work:
  - **Tim Dawborn**
  - **Andrew Naoum**
  - **Glen Pink**

M6: Engineering

- ✗ Battle with Cassandra (to the death?)
- ??? Supported integration with My Masthead
- ✓✓ Refactored NEL to use docrep
- ✓✓ Initiated 4 new developers within a month
- ✓✓ Queue-based parallelism

M6: Research: NEL performance improvement

- TAC 2012
  - ✔✔ Increased performance by 10% points on 2011 data - state-of-the-art
  - ✔✔ 3.1% points off state-of-the art 2012 over kb queries
- SMH dataset
  - ✔✔ Best system performs at 72.44% F-score (61.25% in milestone 3)

M6: Research: Opinions in quotes

- ✔ Presented quote attribution work at EMNLP 2012
- ✔ New annotation tool for marking opinions in quotes
- ✔ Review of existing opinion extraction work
- ✔ Pilot annotation of 700 documents, including some double annotations
- ??? Annotation scheme – has been through several iterations, but is not yet complete

M6: Research: Event linking

- ✔ Presented task description at ACL 2012
- ✔ Built a framework for experimentation
- ✔ Charted and analysed baseline results
- ✔ Used temporal information to constrain event linking queries: try to determine an event's publication date from language
- ✔ Began comparing Fairfax hyperlinks to manual annotations

Filling gaps in our Fairfax archives

Article published in	We had before	We now have
2009	1 051	1 055
2010	396	51 146
2011	57 657	104 559
2012	58 840	86 018

- Five nights so far
- Fetching 1am-6am, waiting 0.2s between downloads
- 261 124 requests ($\approx 50\%$) returned No Content!
- Much better understanding of DCDS assets
    - and their URLs
    - still uncertain about how branding information is stored

Capital Markets CRC Limited

Cassandra

- Goals
  - Move to a faster, more scalable database architecture
  - Reduce the number of databases used in production
- Requires
  - Replacing the existing multiple database architecture with Cassandra
  - Deploying a version of Fizzing Panda that uses Cassandra
- Tasks
  - Design the Cassandra server cluster (FXJ)
  - Design a database schema (CRC)
  - Migrate existing database content (such as the entity store) (CRC)
  - Convert NEL, APIs and the demo to use Cassandra (CRC)

Capital Markets CRC Limited

Progress

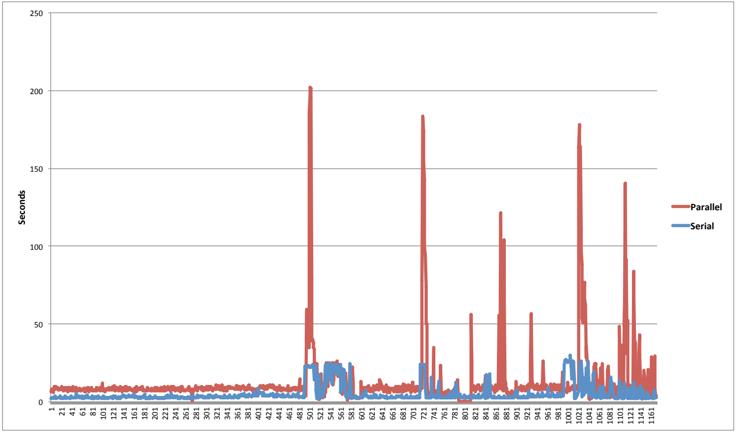
- ✔ An initial database schema has been designed
- ✔ The entity store and data required for linking has been migrated to Cassandra
- ✔ NEL can link end to end using only Cassandra and Solr for full text search
- ✔ Ops have provided a Cassandra cluster for testing
- ✔ Isolated stress tests have been run

Counts 2 Stress Test

Batch Size	1 day of articles (hrs)	1 year of articles (days)
5000	4.04	61.44
2000	1.90	28.87
1000	0.96	14.61
1000 Parallel	2.31	35.2
None	2.42	36.8

- With parallel linkers, it would take over a month to write the counts 2 data for a years worth of articles

Parallel Performance



Gripes

- Counting rows requires reading all counted rows into memory, making counts slow, and placing an upper bound on counts
- Parallel counts on the cluster are affected by spikes in a similar way writes are, making some count operations twice as slow as average
- Mixing reads and writes with multiple clients on one node has caused timeouts and crashes
- At various times Cassandra has lost data, lost indexes, and refused connections with no errors logged

9% improvement on TAC 11 benchmark

System	B^3+ F1
Best TAC 11	**84.6**
CompNews M3	**75.4**
Median TAC 11	71.6
CompNews M6	**84.0**

- puts us in the top-3 systems group in TAC 2011

TAC 2012 Results

System	All	In KB	Not in KB
Best	73.0	**68.7**	84.7
CompNews M6	66.5	**65.6**	67.5
Median	53.6	49.6	59.4

- KB linking is most aligned with CompNews goals
- NIL-clustering will become important for linking the archive
- Within 3.1 points of state-of-the-art for KB linking
- Well above median performance overall (final results to come)

Improving candidate recall

- Early in pipeline and major limiting factor to NEL performance
- Updated Wikipedia: using April 2012 snapshot
- Manually curated statistical rules for name variation
- Boost `title` and `redirect` matches
- In-document coreference for query expansion:
  - Organisational suffixes
  - State abbreviation expansion (e.g. `NSW` → `New South Wales`)
  - Bureaucratic name pre-filtering (e.g. `Dept of Foreign Affairs`)
  - Nicknames (e.g. `Christopher/Chris`)
- Currently 96% recall for TAC 11 data

Statistical Rules for Name Variation

- Given a large sample of entities and their aliases
- Find common transformations
- Over 500,000 rules extracted from Wikipedia
  - `<1> Corporation` → `<1> Corp.`
  - `<1> Inc.` → `<1>`
  - `<1> River` → `River <1>`
  - `<1> University` → `University of <1>`
  - `<1> Texas` → `<1> TX`
  - `United States <1>` → `US <1>`
  - `Bob <1>` → `Robert <1>`

Capital Markets CRC Limited

Supervised NEL

- Gives the model the ability to rank candidates by incorporating many sources of evidence in a principled way
- Very common in TAC 11, we implemented features from systems: dmir_inescid and THU:
  - **String features**: which compute functions between the mention text or its background document, and the candidate title or its background document
  - **Topic features**: the similarity of the topic distributions of the mention and its candidate
  - **NER features**: computed over the sets of named entities detected in the mention document and its candidate document
- Results inconclusive so far, but should allow us to fully exploit the SMH annotated dataset

Unsupervised NEL and NIL thresholding

- Higher search recall means a noisier list (many more candidates)
- Simple average of several metrics
- Minimum score threshold, else NIL
- Tuned on development data

11% performance gain on SMH data in 12 months

Milestone	Mention F1
3	61.25
5	69.80
6	**72.44**

Capital Markets CRC Limited

Areas for improvement. . .

Error	Count	%
Span	930	42.5
Wrong KB	546	24.9
NIL as gold	382	17.4
Gold as NIL	333	15.2

NLP is format/data structure hell

- Many layers of annotation:
  - document metadata (e.g. date, author, categories, . . .)
  - text: sentence boundaries, tokenisation and word normalisation
  - syntax: part-of-speech tags, parse trees
  - semantics: word senses, named entities, argument roles
  - discourse: topics, coreference, entity links
- . . . and headings, paragraphs, hyperlinks, emphasis, etc.
- Each layer has its own legacy text-based format

  ⇒ docs are stored redundantly ⇒ multiple DB queries

  ⇒ custom I/O and nasty, inefficient alignment
- Our tools should have access to all layers and talk about them in the same way

Capital Markets CRC Limited

Holistic NLP with docrep

- An efficient serialisation format for markup and annotations
    - requirements include pointers, references to text spans
    - layers are deserialised only if an app requests them
- Intuitive, declarative API in C++, Python and Java
- One document == one message/blob
- Communicating over streams or sockets
- Local taggers can exploit document-level context

Capital Markets CRC Limited

Tech details: annotation layers in Python

```python
class Token(dr.Ann):
  span = dr.Slice()
  norm = dr.Text()
  pos = dr.Field()

class Mention(dr.Ann):
  span = dr.Slice('Token')
  type_ = dr.Field()
  chain = dr.Pointer('Chain')

class Chain(dr.Ann):
  concept_id = dr.Field()
```

Tech details: document model in Python

```python
class Doc(dr.Doc):
  tokens = dr.Store(Token)
  mentions = dr.Store(Mention)
  chains = dr.Store(Chain)
  asset_id = dr.Field()
  pub_time = dr.DateTime()
```

And accessing documents with annotations:

```python
for doc in dr.Reader(open(filename), Doc):
  for chain in doc.chains:
    print chain, chain.concepts[0]
```

. . . in Java

```
@dr.Doc public class Doc extends AbstractDoc {
  @dr.Store public Store<Token> tokens = new Store<Token>();
  @dr.Store public Store<Mention> mentions = new Store<Mention>();
  @dr.Store public Store<Chain> chains = new Store<Chain>();
}

@dr.Ann public class Token extends AbstractAnn {
  @dr.Field public ByteSlice span;
  @dr.Field public String norm;
}

@dr.Ann public class Mention extends AbstractAnn {
  @dr.Pointer(store="tokens") public Slice<Token> span;
  @dr.Field public String type;
  @dr.Pointer(store="chains") public Chain chain;
}
```

. . . and in C++ (coming soon: Ruby, JavaScript, . . .)

```
class Mention: public dr::Ann {
  dr::Slice<Token *> span;
  std::string type;
  dr::Pointer<Chain> chain;

  class Schema;
};

class Mention::Schema: public dr::Ann::Schema<Mention> {
  DR_POINTER(&Mention::span, &Doc::tokens) span;
  DR_FIELD(&Mention::tag) tag;
  DR_POINTER(&Mention::chain, &Doc::chains) chain;
  Schema(void): dr::Ann::Schema<Mention>("Mention"),
    span(*this, "span", dr::FieldMode::RO),
    type(*this, "type", dr::FieldMode::RO),
    chain(*this, "chain", dr::FieldMode::RW){}
};
```

Capital Markets CRC Limited

Tech details: the wire format

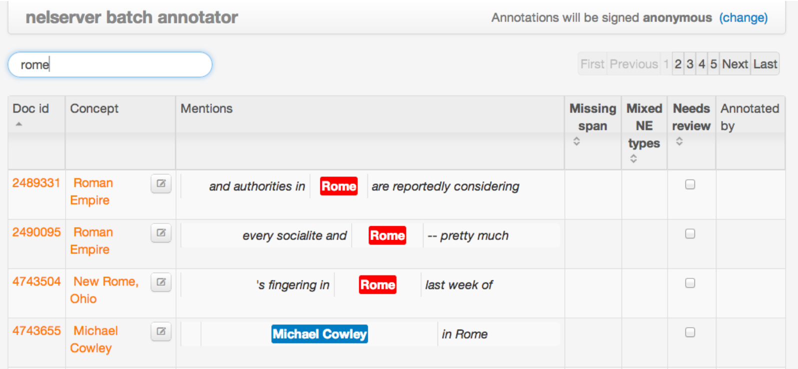
- Based on msgpack
  - A JSON-like binary format
  - But small and fast (before compression!)
  - e.g. each of the following takes 1 byte:
    - integer $-32 \leq n < 128$, true, false, null
    - header for map or array of length < 16
    - header for string of length < 32
- On top of msgpack, docrep:
  - uses small integers where possible
  - but avoids unnecessary data obfuscation
  - stores the data model with each doc ⇒ self-describing
  - (un)swizzles pointers
  - deserialises annotation stores only when needed
  - provides lazy dynamic access to other annotations

Implementation status for Computable News

- ✔ Import and tokenisation of RLAY stories
- ✔ Import and tokenisation of syndication feeds
- ✔ C&C tools: POS tags, parse trees, named entities
- ✔ Entity and concept linking
- ✔ Quotation extraction and attribution
- ✔ Event linking experimentation
- ✔ Story rendering to HTML
- ⇒ fewer custom formats
- ⇒ eliminated a large and confusing DB table
- ⇒ more consistent internal code structure
- ⇒ consistent streaming and socket interfaces

Batch correction interface

- Editors can override linking decisions for display
- Annotators can clean up errors, yielding higher-quality data



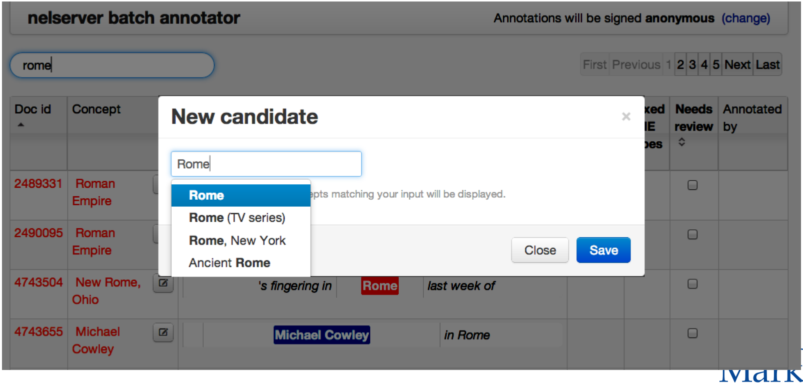
nelserver batch annotator Annotations will be signed **anonymous** (change)

rome

Doc id	Concept	Mentions	Missing span	Mixed NE types	Needs review	Annotated by
2489331	Roman Empire	*and authorities in* **Rome** *are reportedly considering*			☐	
2490095	Roman Empire	*every socialite and* **Rome** *-- pretty much*			☐	
4743504	New Rome, Ohio	*'s fingering in* **Rome** *last week of*			☐	
4743655	Michael Cowley	**Michael Cowley** *in Rome*			☐	

Batch correction interface

- Editors can override linking decisions for display
- Annotators can clean up errors, yielding higher-quality data



nelserver batch annotator

Annotations will be signed **anonymous** (change)

rome

First Previous 1 2 3 4 5 Next Last

Doc id | Concept | …xed …IE …pes | Needs review | Annotated by

2489331 Roman Empire □

2490095 Roman Empire □

4743504 New Rome, Ohio — 's fingering in **Rome** last week of □

4743655 Michael Cowley — **Michael Cowley** in Rome □

New candidate ×

Rome

…pts matching your input will be displayed.

Identifying opinions in quotes

- We have quotes, topics, and entities
- Can we identify the opinion an entity holds about a topic, via the quotes they're making?

Mr Abbott says that given that change is habit-forming, the best way to stop the "republican extremism" that the conservative republicans say they fear is "to vote down any republic at all – in the same way that deterring major crime requires tackling petty crimes"

Task description

- Identify documents on a topic of interest via keyword search
- Define a "topic statement" which sets out a position on the topic
- Classify quotes as "supporting", "neutral", or "opposing"

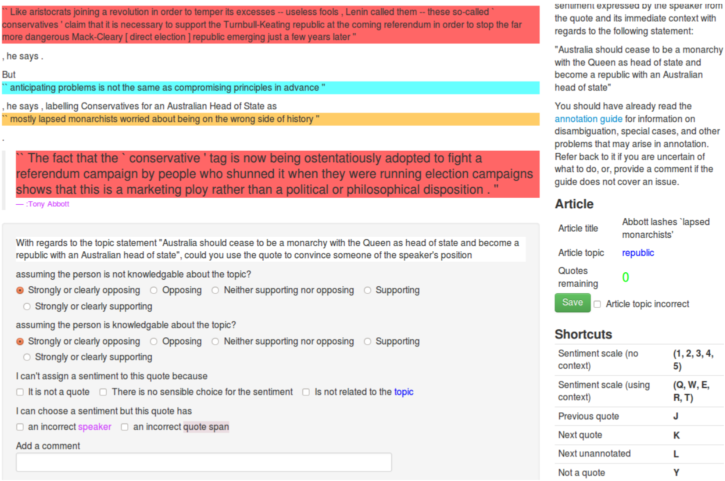
Progress

- ✓ Annotation tool
- ✓ Review of existing technologies
- ✓ Pilot annotation of 700 documents over 7 topics
  - Republic
  - Carbon tax
  - Reconciliation
  - Work choices
  - Abortion
  - Immigration
  - Same sex marriage

??? Annotation scheme – has been through several iterations

Capital Markets CRC Limited

Annotation tool



`` Like aristocrats joining a revolution in order to temper its excesses -- useless fools , Lenin called them -- these so-called ` conservatives ' claim that it is necessary to support the Turnbull-Keating republic at the coming referendum in order to stop the far more dangerous Mack-Cleary [direct election] republic emerging just a few years later "

, he says .

But
`` anticipating problems is not the same as compromising principles in advance "

, he says , labelling Conservatives for an Australian Head of State as
`` mostly lapsed monarchists worried about being on the wrong side of history "

.

> `` The fact that the ` conservative ' tag is now being ostentatiously adopted to fight a referendum campaign by people who shunned it when they were running election campaigns shows that this is a marketing ploy rather than a political or philosophical disposition . "
>
> — :Tony Abbott

With regards to the topic statement "Australia should cease to be a monarchy with the Queen as head of state and become a republic with an Australian head of state", 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

Add a comment

sentiment expressed by the speaker from the quote and its immediate context with regards to the following statement:

"Australia should cease to be a monarchy with the Queen as head of state and become a republic with an Australian head of state"

You should have already read the annotation guide for information on disambiguation, special cases, and other problems that may arise in annotation. Refer back to it if you are uncertain of what to do, or, provide a comment if the guide does not cover an issue.

Article

Article title	Abbott lashes `lapsed monarchists'
Article topic	republic
Quotes remaining	0

Save □ Article topic incorrect

Shortcuts

Sentiment scale (no context)	**(1, 2, 3, 4, 5)**
Sentiment scale (using context)	**(Q, W, E, R, T)**
Previous quote	**J**
Next quote	**K**
Next unannotated	**L**
Not a quote	**Y**

Named Entity Linking

- Write TAC 12 system report
- Supervised NEL
- Apposition
- Batch annotation to improve training data quality
- Batch correction to identify specific problem areas
- next steps in live linking editor

NER in noisy text

- Prerequisite for NEL
- Existing algorithms can be fragile when given noisy text
- **CompNews applications**:
  - **ocr**: scanned historical archives, books
  - **transcripts**: video, audio
  - **social text**: comments, blogs
  - **translations**: transliterated names

Slot-filling

- Find the values of specified attributes of an entity (“slots”) in a collection of documents
  - e.g. for a person find their title, age, country of residence, employer, family, etc.
- Currently surveying current approaches and setting up a baseline system for slot-filling
- **CompNews applications**:
  - **nel**: one person cannot have two (correct) dates-of-birth!
  - **editor**: extracting information to insert into news text