The Computable News project: research in the newsroom

Will Radford [*]
Xerox Research Centre Europe
6 chemin de Maupertuis
38240 Meylan, France
will.radford@xrce.xerox.com

Daniel Tse [†]
School of Information Technologies
University of Sydney
NSW 2006, Australia
daniel@overpunch.com

Joel Nothman
School of Information Technologies
University of Sydney
NSW 2006, Australia
joel@it.usyd.edu.au

Ben Hachey [‡]
School of Information Technologies
University of Sydney
NSW 2006, Australia
ben.hachey@sydney.edu.au

George Wright
Fairfax Media
1 Darling Island Road
NSW 2009, Australia
gwright@fairfaxmedia.com.au

James R. Curran
School of Information Technologies
University of Sydney
NSW 2006, Australia
james@it.usyd.edu.au

ABSTRACT

We report on a four year academic research project to build a natural language processing platform in support of a large media company. The *Computable News* platform processes news stories, producing a layer of structured data that can be used to build rich applications. We describe the underlying platform and the research tasks that we explored building it. The platform supports a wide range of prototype applications designed to support different newsroom functions. We hope that this qualitative review provides some insight into the challenges involved in this type of project.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing

Keywords

Named entity linking; Online news applications; Event linking; Quotation extraction and attribution

1. INTRODUCTION

Traditional news media faces increasing commercial pressure as advertising revenue streams decline and it competes with natively-digital new media for readers' attention. The need to generate more news, but in less time and with fewer resources, has encouraged companies to innovate with the aim of building compelling news products. It can be difficult for companies to justify full-time research departments or divert existing staff to high-risk research activities.

We report on a joint project between a media company, a university research lab and a commercialisation-oriented research centre that ran from 2010 to 2013. Fairfax Media is a major publisher of newspapers, magazines, radio and websites in Australia and New Zealand. For example, their largest site, the Sydney Morning Herald, has an Alexa ranking of 17 in Australia and 1,029 globally, the second largest news website in Australia behind news.com.au, a News Corp Australia site ranked 15 in Australia.[1] Moreover, Fairfax has an extensive network of regional radio, newspapers and websites. Schwa Lab is a natural language processing (NLP) research group at the University of Sydney and works in wide range of areas. The Capital Markets Cooperative Research Centre works within an Australian Federal Government funding regime that promotes research partnerships between industry partners and academic institutions with a view to training PhD students and producing commercializable research. This collaborative structure has two main advantages. Researchers have access to large-scale commercial data and use cases for applications, while industry partners can essentially outsource research to experienced teams less affected by daily commercial pressures.

The paper is structured as follows: we first motivate and describe the *Computable News* platform. This includes discussion of the three research areas that satisfied the scientific agenda for the project: named entity linking, quotation extraction and attribution and event linking. Next we detail the prototype applications that use the platform and are designed to show how NLP can support news media, including tools for journalists, entity-oriented archive browsing, summarisation, news games, editorial analytics and knowledge base maintenance. We hope that this qualitative review gives some insight for anyone seeking to apply academic research in a news media innovation context.

[*]Work done while the author was at the University of Sydney and the CMCRC.

[†]Now at Lumific.

[‡]Work done while the author was at Macquarie University and the CMCRC.

[1]`http://www.alexa.com` rankings measured on 21/1/2015.

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2742009.

2. THE COMPUTABLE NEWS PLATFORM

The vision for the platform is that data is critical for modern news publishers, but it is difficult to extract from their main holdings – vast collections of unstructured text. We wish to interpret news stories, producing structured data that can be used by downstream applications. The developers of these applications are already media experts: data scientists, journalists and developers, and should not have to become NLP experts to use data extracted from text. The platform model promotes data reuse and quicker integration of facts and related content into stories.

The project required a deliberate research strategy to explore three areas that provided scientific challenges and could help drive the platform. The first was named entity linking (NEL), which was the central to the project and featured most in the platform. This helped enable quotation extraction work, as this required entities to which quotes could be attributed. Finally, event extraction was an interesting task, but one that presented substantial scientific challenges.

2.1 Named entity linking

Named entity linking is the task of linking names in text to a knowledge base (KB). For example, given an ambiguous name George Bush, a system must select the correct KB record (we use Wikipedia pages). The task also makes the realistic assumption that no KB is ever complete, so some mentions of George Bush do not refer to a former US president and are new, NIL entities. Moreover, there may be multiple distinct NILs named George Bush, and clustering them correctly is necessary if we want to automatically build and curate KBs from text.

Disambiguating names in news text is fundamental to the COMPNEWS platform as it enables an *entity-centric* view of the news, where readers can access stories through the people, organisations and places that drive them. NEL can be seen as a process of *extraction*, *search* and *disambiguation*, and the approaches used in our platform follow this model. Extraction consists of named entity recognition followed by in-document name coreference. We use the `candc` maximum entropy sequence tagger[2] trained on a corpus of annotated SMH stories (see below for details), and a series of high-precision heuristics to match name mentions in the document.[3] We then identify the longest, and hopefully most informative, mention in each coreference chain and search for candidate entity matches in our KB– a snapshot of Wikipedia.[4] We use a full-text index populated with alternative names mined from the articles of Wikipedia entities. This presents a trade-off, as overly precise search will miss non-standard spellings of entity names, but too many candidates introduces scaling and noise problems for disambiguation. Our disambiguation methods use *whole document* information by using features from other candidates entities in the document to help rerank a mention's candidate list so that the best candidate link is ranked first. Finally, we model the NIL case by considering mentions without entity candidate above a threshold to be linked to NIL. We experimented with directly supervised models for linking and these performed the best. We also developed a simple system that combines several features from the literature including: similarity between parts of the document and entity article text, KB statistics for entity popularity and name suitability[5], and metrics that exploit the Wikipedia article graph and categories. This unsupervised system proved a strong, robust baseline and was used in our platform, as well as in shared task submissions [14].

Linked entities become story metadata, allowing recommendation and smarter search that can present different candidates for an ambiguous query. The platform also treats entity co-occurrence in the same story as indicative of a relation between the entities, so the platform calculates and stores these statistics. In addition to linking news stories, we also link image captions, which allows us to associate images with entities, and automatically create a entity image galleries. Since it was the core of the platform, building an accurate, robust and fast NEL system was critical.

We relied on a few components and assumptions to help scale the platform to large news archives. The first is that linking a document, beyond access to the KB, can be performed in isolation and so parallelised. We were operating over around 20 years of news stories and this meant that the entire archive could be linked overnight with a small cluster of servers. To provide fast access to the KB, we used a range of database technologies: Apache Solr[6] for full-text search over entity names and linked stories and PostgreSQL[7] to support our annotation servers. The storage of co-occurrence statistics is challenging as, for a document with n entities, the number of relations is $O(n^2)$.[8] This performed reasonably well in practice as there are few long documents with many entities, although sporting match previews and reviews with full team lists can be problematic. We experimented with Cassandra[9], but had problems scaling our instances to store incoming counts from multiple linking processes in parallel, so used Hypertable[10] to store count and KB data. Finally, we used a dedicated document representation framework, DOCREP [1], to store stories and any annotation over them. The framework is designed for efficient storage and object modelling of documents, using standoff annotation that we used to represent labelled text spans such as entities, events and quotations. Additionally, it follows UNIX streaming patterns and has a comprehensive suite of tools for managing large collections of text [6].

2.2 Quotation extraction

Accurately reporting what is said, who said it and when is central to journalism. Publications differ in the extent to which they use direct quotation (with punctuation) or indirect quotation or paraphrases, often using the latter to represent events in a more sensational manner. Increasingly, what political figures can be shown to have said, or not said, is the focus of much journalistic and political effort, so accurate extraction and attribution is important to a news platform. From the readers' point of view, quotes are the distillation of the speaker's position on an issue and are thus eminently shareable on social channels. A large col-

[2] `http://svn.ask.it.usyd.edu.au/trac/candc`

[3] We only consider proper names, not pronouns or common nouns.

[4] `http://www.wikipedia.org`

[5] Specifically, p(entity|mention), e.g., p(United_Nations|UN)

[6] `http://lucene.apache.org/solr`

[7] `http://www.postgresql.org`

[8] We store counts for entities `(a,b)` and `(b,a)`, which allows faster access through key prefixes at the extra cost of storage.

[9] `http://cassandra.apache.org`

[10] `http://hypertable.org`

Figure 1: Screenshot of the editor. The left panel contains the story markup with hyperlinked entity names. The screenshot has been taken immediately after the system had replaced the macro `{partner}` with Melissa Babbage, a value extracted from Wikipedia. The middle panel shows the rendered story and the right hand panel shows information about the linked entities.

lection of attributed quotes would allow journalists to check consistency on issues over time, monitor the dissemination of quotes over social media and gauge the impact of stories, and work in support of journalistic integrity auditing to minimise libel complaints.

We explored statistical methods for extracting quotations from text and attributing them to identified entities [10]. This work evaluated several systems for extracting direct quotes on SMH, Wall Street Journal (WSJ) and literature corpora. The systems included a sequence model for quote attribution, that performed well on the WSJ dataset, and a rule-based baseline. This attributes quotes to the closest entity in previous sentences if a reported speech verb is found, else the first entity after the quote. Despite its simplicity, this performed well on the SMH data (91.2% accuracy) and was used where we required quotation extraction in the platform. Once quotations are extracted and attributed, they can be made available as part of the platform information about their entity. We also investigated how quotations related to positions on debate issues, annotating a corpus of opinions. However, it is sometimes difficult for humans to gauge whether a quote is in favour or opposition to a debating position.

2.3 Event linking

The project also attempted to characterise events, but this is a difficult task in practice. Entities are fairly well-understood targets for linking as linking to entities requires simply choosing between candidates and NIL. Events are far harder to characterise and suffer from problems of identity; it can be sometimes hard to decide when one event ends and another begins. These are also differently understood in different domains: micro-level events such as mergers and share issues may be useful for financial news, but larger amorphous events such as a government scandal may be of more interest to general news. We framed the problem of event linking as that of automatically hyperlinking descriptions of events in news to a story in the archive that describes it [7]. Event linking was not eventually transferred to the platform, but we saw it as potentially useful to help journalists write richer linked stories more easily, and also to automatically add structure to large digital archives that predate manual hyperlinking.

2.4 Research contributions

Maintaining a strong research output was important to the scientific aspect of the project, either reporting directly on project techniques or applying them to different datasets, as was the case with quotation extraction. Project members made research contributions in these areas: named entity linking [4, 3, 2, 13] and apposition extraction [15], event linking [7, 5] and quotation attribution and extraction [10, 9, 11, 8]. We also participated in a shared task to evaluate NEL at the Text Analysis Conference (TAC)[17, 16, 14, 12]. This allowed us to compare to research state-of-the-art on common dataset and consistently good results were a positive story for promoting the project inside Fairfax. Finally, all three directions required manual annotation of data for evaluation and system training, and the design of schemes and tools was a large component of our research work. Our annotation tasks are moderately complex and require annotator training, so we opted for Freelancer.com[11], a crowdsourcing market that is more suitable for training and retaining annotators than other volume-oriented markets. In conclusion, the platform helped us satisfy the research goals of the project, in the next section, we present the applications that we built on top of it – the main way in which we engaged journalists and developers at Fairfax.

3. APPLICATION PROTOTYPES

Although our main task was to research and implement the platform, it soon became apparent that in order to maintain the project and the relationship with industry partners, these tools needed to be "sold" through working demonstration applications using the partner's data. In this section, we present the applications that we built, in the general order that they apply to the news lifecycle.

3.1 Editor: supporting journalists

We want to enable journalists to generate rich stories as quickly and seamlessly as possible, embedding video and images, lists of related stories and facts and statistics. Figure 1 shows a screenshot from the prototype we developed to demonstrate how this might work. It shows three pan-

[11] `https://www.freelancer.com`

Figure 2: Screenshot of Zoom for the mining company BHP Billiton. The top half shows a timeline of related stories in a carousel. The bottom half shows a "doughnut" of the eight most related entities in the KB, and links to recent stories.

els: story draft in `markdown`[12], rendered story and the entity stack.[13] As the journalist types their story, the platform links entity mentions in text and inserts hyperlinks into the text and populates the entity stack on the right side of the screen. This provides implicit cues to the journalist, as if the platform fails to link a particular mention, it may indicate a misspelling or non-standard way of referring to an entity.[14] Another benefit of linking entities in text is that it supports a system of macros that can take advantage of structured information from the KB. For example, the screenshot is taken immediately after the system has replaced the macro `{partner}` with the name of Joe Hockey's partner: Melissa Babbage. This uses some proximity heuristics to determine which entity to look up and provides a framework for information integration into the story. Other macros were defined for currency conversion using web APIs, date-of-birth, current age, stock tickers for live instrument pricing, and static map display using location coordinates and web mapping APIs. The final potential benefit is that it allows and motivates journalists to provide feedback to correct the system as they are writing. Busy journalists are unlikely to spend time annotating stories or manually adding tags or links, so it is important to provide a good incentive for any extra work they may perform. The feedback proposition for the smart editor is that it: adds to training data that will improve the NEL system, allows the system to add facts and related media that may otherwise have to be manually added, saving the journalist time.

[12] `http://daringfireball.net/projects/markdown`
[13] A production version of this tool would almost certainly need to exchange the markup for a WYSIWYG interface.
[14] Although this may be done for comedic effect.

3.2 Zoom: entity-oriented news

The main product built on the platform was an entity-oriented news product named zoom. The motivation behind this was to automatically create a landing-page for prominent entities in the news. Figure 2 show the entity page for a large mining company, BHP Billiton. The top section is a timeline of stories mentioning the company linked to a gallery showing the headline, story snippet and image if available. Users can navigate using the image carousel or the timeline and click through to the story's web page to read it in full. It is not feasible to show all stories for very popular entities, so we use a supervised linear regression model to rank stories for a time period to identify the most useful stories to show on the timeline. The bottom half of the page shows a "doughnut" of the eight most related entities in the KB: other mining companies, stock indices and mining regions. The tail of the page contains recent stories that mention the company along with thumbnail images. Zoom was launched in early 2013 at the same time as a paywall as a benefit to subscribers. The development process was a good opportunity to engage with the Fairfax developers building the frontend, and explain how the platform worked. An interesting aspect of the timeline archive view is that it can drive traffic to historical stories. These stories are rarely featured on the front page of the website and hence unlikely to be visited by users. One design goal of zoom was to allow users to explore the archive through entities and stumble on old stories. As well as the entity pages, stories that had been linked featured entity tags that would link through to the entity pages. Coupled with user tracking, this presents a way of building user profiles not of what users read, but of the entities they follow.

3.3 Shorty: summarization

We wished to explore how NEL, quotation extraction and coreference can be exploited to perform extractive summarization. Taking NEL output from the platform with coreference chains as the starting point, we created a web interface which, given the text of a news story, allowed a human editor to create a shorter précis of the story, by specifying a *budget* which controls the proportion of sentences retained. This has applications when screen real estate is limited, such as for mobile and wearable devices, and for any case where it is important for the reader to absorb the story as quickly as possible.

The editor selects the budget using a slider control whose two extremes are labelled MORE (selecting all of the sentences) and LESS (which selects only one sentence). Intermediate values of the slider then select intermediate numbers of sentences. Given a budget, a scoring algorithm determines which full sentences from the source story to include in the précis. For each sentence, a relevance score is calculated, and the k sentences with the greatest score are included (where k is determined by the LESS...MORE slider). The score is a weighted sum of normalised values corresponding to the criteria in Table 1. The weights are specified by the editor and

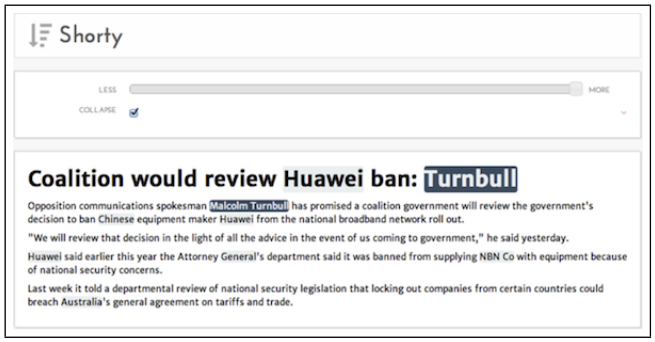

Figure 3: Screenshot of shorty. As the user drags the slider from MORE to LESS, sentences will fade out, showing a summary of the text. An advanced interface allows weighting the different factors that govern how important sentences are in the summary. The selected entities Turnbull and Malcolm Turnbull are coreferent, and this allows us to rewrite the former to Malcolm Turnbull for fluency.

Criterion	*Meaning*
1ST SENTENCE	1 for first sentence, 0 otherwise
HEADLINE	1 for first sentence, 0 otherwise
QUANTITATIVE	Proportion of numeric tokens
OPINIONATED	Main verb is opinion-expressing
IMPORTANT	Max entity prior score over entities
RELEVANCE	Max relevance score over entities
QUOTES	Presence of direct quotes
SENTIMENT	Max absolute value of sentiment
VAGUE	Presence of passive constructions

Table 1: Criteria for displaying sentences in shorty.

allow the editor to express the likelihood that a sentence is selected for the précis, based on its syntactic, semantic and orthographic characteristics.

We observed that simply extracting sentences from the full story could cause readability issues if we removed the sentence with the canonical form of a named entity, but kept an ambiguous mention. We exploited NEL and coreference in our system to solve the problem. Using the coreference information, we replace the first mention of every named entity with its full (canonical) form. This transformation maintains the readability of the resulting précis, despite the fact that the first mention of a given named entity may not have been selected for the précis.

Our summarization method is fairly simple and is extractive as it chooses sentences to combine into an extract rather than generate new text (i.e. abstractive). While this may be less sophisticated, we consider it a low-risk, robust method, especially when combined with coreference repair.

3.4 Amp: news trivia

Most applications we considered were serious products with a business purpose, but we also tried to think of less conventional ways to use the platform. Amp[15] is a news trivia game that uses platform information about people in the news. Figure 4 shows a screenshot of the game, with four first names on screen. The top name is in "focus" and the

[15]For ampersand.

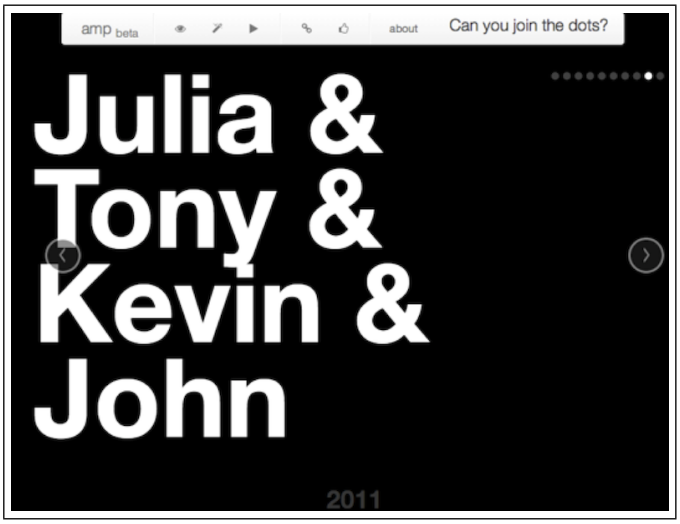

Figure 4: Screenshot of amp. This displays the first names of newsworthy people who are related to one another – the user must guess who they are. The user can reveal the surnames of the people, click through to "focus" on another entity or look how this changes over time in the carousel.

task is to guess the identity of the other three related people in that year. The user can choose to reveal the surnames and look at the top related people for the focus entity in other years. Alternatively they can click on a name to guess again with a new focus entity.

Although lighthearted, amp contains some design features that take advantage of the platform. Since it is based on the linked news archive continually updated with new stories, it is difficult to predict the people who feature in the game. This can be positive, as the game is pleasantly surprising, but can also risk insensitivity when displaying the names of recently deceased people, although it provides an interesting entity-centric obituary focussing on related people over time.

3.5 Insight: supporting editorial

Analytics is becoming increasingly important in news media, so another application switched focus from monitoring entities that were written about to keeping track of what entities were read about and shared on social media. This information was updated automatically and fed a dashboard that visualised trending entities over time. The dashboard was split into pageview and Twitter share tracking, and sharing or viewing a story is used as a proxy for applying those actions to the entities it contains. Although we concentrated on Fairfax publications, the dashboards can be targeted at other news sites or aggregators to produce an early warning system of sorts to give editorial oversight over trending entities.

3.6 Correction: maintaining KBs

The final application considers the difficult task of KB maintenance and curation. While web-scale KBs are useful and underpin the platform, they are simply too large to check or fix exhaustively. This motivates a set of quality assurance tools for correcting the KB and decisions made by the NEL system. The correction tool allows a user to check and review the links assigned by the system, and records

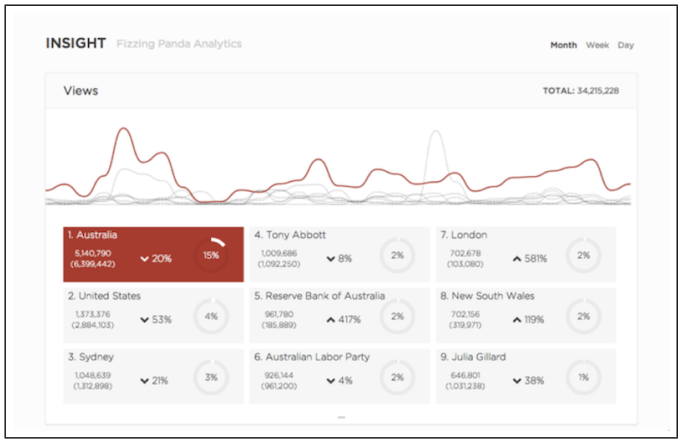

Figure 5: Screenshot of insight. This shows the most popular concepts by virtue of their page views in SMH articles as well as trend information. Another view shows similar information, but considering article shares on Twitter.

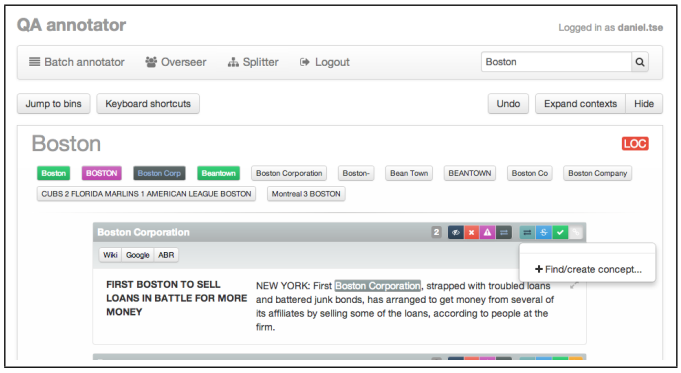

Figure 6: Screenshot of the correction tool. This screen focusses on mentions linked to the US city of Boston. Mentions are displayed as tags, some correct Boston and BEANTOWN, some incorrectly linked Boston Corp. and some incorrectly identified mentions Montreal 3 BOSTON. The user can browse the links in context and make individual or bulk changes to the KB.

decisions to feedback to improve system performance. For example, Figure 6 shows a range of mentions that have been linked, some incorrectly, to the US city of Boston. As well as fixing errors, it is important to be able to add new entities for the KB, perhaps importing from Wikipedia. This is critical for breaking stories, for example Edward Snowden did not have a KB entry before he became a story and thus could not be linked to. Applying these tools in a business context also have interesting side effects. One prominent source of errors is common in local or regional news contexts, where someone prominent in a small community who happens to share a name with a newsworthy KB entity may be mistakenly linked to the KB, which suggests that a concept of localisation is required for KBs and linking systems. Others effects are cultural in that academics are used to considering evaluations over fixed data that yield a performance score. These datasets and metrics may be designed to test challenging research cases, and so performance of 80% is not an unreasonable accuracy, for example. In a commercial setting, this translates to one in five cases wrong, which may not be acceptable. Another problem is related to different error types. The usual testing paradigm can tell you which examples are correct and which are not. It tells you little about *how bad* the error was, and we found several cases of easy-to-fix but egregious errors (e.g. linking the Victorian cricket team to "Queen Victoria" as Victorian is often used in Wikipedia in reference to her reign). Although difficult to develop and test, these systems force researchers and engineers to face difficult questions that must be answered to operate large and dynamic KBs.

3.7 Reception

We are unable to release sensitive quantitative information about the impact of COMPNEWS on Fairfax business, but can share anecdotal evidence about the reception of users, journalists and product managers.

Participants in Zoom user studies rated it highly for topic-driven news consumption, for example users seeking more context on stories they had not been following or had related older content from the archive. This supports the use-case for deeper dives into the archive and, indeed, participants rated this view as less useful for browsing daily news. Journalists were involved in the project from the beginning, but only became enthusiastic when demonstrations showed how it could leverage older content and suggest related stories. They also saw potential as an exploratory tool for examining the links between people mentioned in large collections such as official government Hansards and documents requested through freedom of information processes. Journalists are increasingly engaged in making their content discoverable by social media and search engines, so any tools that help automate the process were seen as positive. Finally, news product managers were most excited in recommending readers more content to keep them engaged and consuming. This was seen as a way to help differentiate the products from the many others vying for the readers' attention.

4. CONCLUSION

This paper reports on a four year joint project that explored how academic research could be applied to support a commercial news media business. We presented the COMPNEWS platform that was the vehicle for research contributions in named entity linking, quotation extraction and attribution, and event linking. Moreover, it allowed us to build prototype applications that demonstrated how NLP technology could be used to support the different stages of a news story: writing, displaying, promoting and analytics. We hope that this qualitative review provides some insight into the advantages and challenges of this type of endeavour.

5. ACKNOWLEDGEMENTS

Thanks must also go to the other members of Schwa Lab (especially Tim Dawborn), our Freelancer annotators, and supporters at Fairfax and the CMCRC. This work was supported by ARC Discovery grant DP1097291.

6. ADDITIONAL AUTHORS

Additional authors who worked on the project while at the University of Sydney and the CMCRC: Will Cannings, Tim O'Keefe, Matt Honnibal, David Vadas and Candice Loxley.

7. REFERENCES

[1] T. Dawborn and J. R. Curran. docrep: A lightweight and efficient document representation framework. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 762–771, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics.

[2] B. Hachey, J. Nothman, and W. Radford. Cheap and easy entity evaluation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 464–469, Baltimore, Maryland, June 2014.

[3] B. Hachey, W. Radford, and J. R. Curran. Graph-based named entity linking with Wikipedia. In *Proceedings of the 12th International Conference on Web Information System Engineering*, Sydney, NSW Australia, 2011. Springer.

[4] B. Hachey, W. Radford, J. Nothman, M. Honnibal, and J. R. Curran. Evaluating entity linking with Wikipedia. *Artificial Intelligence*, 194:130–150, January 2013.

[5] J. Nothman. *Grounding event references in news*. PhD thesis, School of Information Technologies, University of Sydney, Sydney, Australia, 2014.

[6] J. Nothman, T. Dawborn, and J. R. Curran. Command-line utilities for managing and exploring annotated corpora. In *Proceedings of the Workshop on Open Infrastructures and Analysis Frameworks for Human Language Technologies*, Dublin, Ireland, August 2014.

[7] J. Nothman, M. Honnibal, B. Hachey, and J. R. Curran. Event linking: grounding event reference in a news archive. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 228–232, Jeju, Korea, July 2012.

[8] T. O'Keefe. *Extracting and Attributing Quotes in Text and Assessing them as Opinions*. PhD thesis, School of Information Technologies, University of Sydney, Sydney, Australia, 2014.

[9] T. O'Keefe, J. R. Curran, P. Ashwell, and I. Koprinska. An annotated corpus of quoted opinions in news articles. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 516–520, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[10] T. O'Keefe, S. Pareti, J. R. Curran, I. Koprinska, and M. Honnibal. A sequence labelling approach to quote attribution. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 790–799, Jeju, Korea, July 2012.

[11] S. Pareti, T. O'Keefe, I. Konstas, J. R. Curran, and I. Koprinska. Automatically detecting and attributing indirect quotations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013.

[12] G. Pink, W. Radford, W. Cannings, A. Naoum, J. Nothman, D. Tse, and J. R. Curran. SYDNEY-CMCRC at TAC 2013. In *Proceedings of the Text Analysis Conference*, Gaithersburg, MD USA, November 2013. National Institute of Standards and Technology.

[13] W. Radford. *Linking Named Entities to Wikipedia*. PhD thesis, School of Information Technologies, University of Sydney, Sydney, Australia, 2015.

[14] W. Radford, W. Cannings, A. Naoum, J. Nothman, G. Pink, D. Tse, and J. R. Curran. (Almost) Total Recall – SYDNEY-CMCRC at TAC 2012. In *Proceedings of the Text Analysis Conference*, Gaithersburg, MD USA, November 2012. National Institute of Standards and Technology.

[15] W. Radford and J. R. Curran. Joint apposition extraction with syntactic and semantic constraints. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 671–677, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[16] W. Radford, B. Hachey, M. Honnibal, J. Nothman, and J. R. Curran. Naive but effective NIL clustering baselines – CMCRC at TAC 2011. In *Proceedings of the Text Analysis Conference*, Gaithersburg, MD USA, November 2011. National Institute of Standards and Technology.

[17] W. Radford, B. Hachey, J. Nothman, M. Honnibal, and J. R. Curran. Document-level entity linking: CMCRC at TAC 2010. In *Proceedings of the Text Analysis Conference*, Gaithersburg, MD USA, November 2010. National Institute of Standards and Technology.