Finding Names in Trove: Named Entity Recognition for Australian Historical Newspapers

Sunghwan Mac Kim*
Data61, CSIRO,
Sydney, Australia
Mac.Kim@csiro.au

Steve Cassidy
Department of Computing
Macquarie University
Sydney, Australia
Steve.Cassidy@mq.edu.au

Abstract

Historical newspapers are an important resource in humanities research, providing the source materials about people and places in historical context. The Trove collection in the National Library of Australia holds a large collection of digitised newspapers dating back to 1803. This paper reports on some work to apply named-entity recognition (NER) to data from Trove with the aim of supplying useful data to Humanities researchers using the HuNI Virtual Laboratory. We present an evaluation of the Stanford NER system on this data and discuss the issues raised when applying NER to the 155 million articles in the Trove archive. We then present some analysis of the results including a version published as Linked Data and an exploration of clustering the mentions of certain names in the archive to try to identify individuals.

1 Introduction

In recent years, digitised newspaper archives have appeared on the web; they make fascinating reading but also provide important primary sources for historical research. The Trove (Holley, 2010)[1] Newspaper collection at the National Library of Australia (NLA) provides an interface for users to search and browse the collections of scanned pages using an optical character recognition (OCR) based transcript of each article. While the OCR results contain errors, they provide enough detail to enable a full-text index to return relevant results to search terms. The documents stored in the Trove archive are made freely available for any purpose by the National Library of Australia.

An abundance of natural language processing (NLP) tools have been developed for Digital Humanities (Brooke et al., 2015; Scrivner and Kübler, 2015) and such tools can greatly facilitate the work of Humanities scholars by automatically extracting information relevant to their particular needs from large volumes of historical texts. This project explores the use of Named Entity Recognition on the Trove Newspaper text to provide a resource for Humanities scholars.

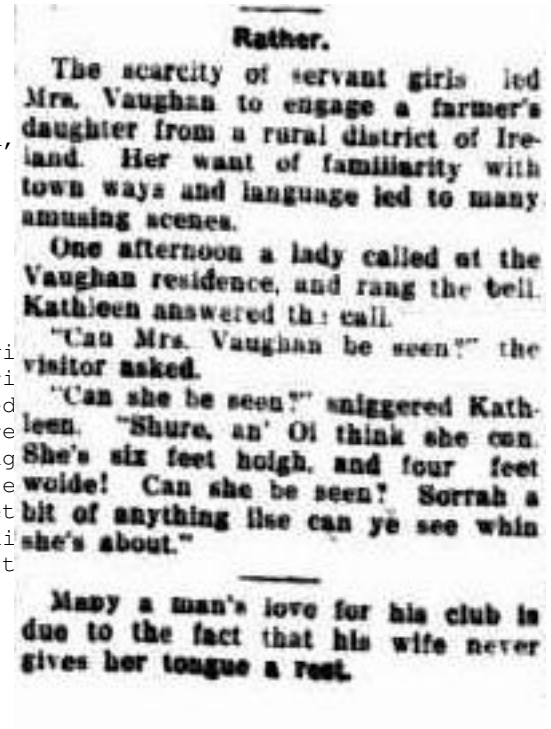
Newspapers are an important repository for historical research. Digitisation of newspaper text via Optical Character Recognition (OCR) enhances access and allows full text search in the archive. It also supports more sophisticated document processing using Natural Language Processing (NLP) techniques. Europe and the United States have actively participated in research on digitised historical newspapers and developed web-based applications using NLP to provide visualisation of useful information (Willems and Atanassova, 2015; Torget et al., 2011). The web-based applications have empowered digital humanities scholars to efficiently exploit historical newspaper content. The Europeana Newspapers project was performed to provide access to digitised historical newspapers from 23 European libraries (Willems and Atanassova, 2015). They used 10 million newspaper articles produced by OCR and a number of tools were developed for researchers. In particular, named entity recognition (NER) was applied to extract names of persons, places and organisations from the digitised newspapers. The University of North Texas and Stanford University used NER and topic modelling on 1 million digitised newspaper articles (Torget et al., 2011). They built interactive visualisation tools to provide researchers with the ability to find language patterns

The work was done while the first author was a research associate at Macquarie University and was supported by a grant from NeCTAR.

[1] http://trove.nla.gov.au/

Sunghwan Mac Kim and Steve Cassidy. 2015. Finding Names in Trove: Named Entity Recognition for Australian Historical Newspapers . In *Proceedings of Australasian Language Technology Association Workshop*, pages 57−65.

```
{
 "id":"64154501",
 "titleId":"131",
 "titleName":"The Broadford Courier (Broadford,
 "date":"1917-02-02",
 "firstPageId":"6187953",
 "firstPageSeq":"4",
 "category":"Article",
 "state":["Victoria"],
 "has":[],
 "heading":"Rather.",
 "fulltext":"Rather. The scarcity of servant gi
 engage a farmer's daughter from a rural distri
 of familiarity with town ways and language led
  One afternoon a lady called at the Vaughan re
   Kathleen answered the call.' \"Can Mrs. Vaug
   asked. \"Can she be seen?\" sniggered Kathle
   she can. She's six feet hoigh, and four feet
   Sorrah a bit of anything ilse can ye see whi
   man's love for his club is due to the fact t
   gives her tongue a rest",
 "wordCount":118,
 "illustrated":false
 }
```

Rather.

The scarcity of servant girls led Mrs. Vaughan to engage a farmer's daughter from a rural district of Ireland. Her want of familiarity with town ways and language led to many amusing scenes.

One afternoon a lady called at the Vaughan residence, and rang the bell. Kathleen answered the call.

"Can Mrs. Vaughan be seen?" the visitor asked.

"Can she be seen?" sniggered Kathleen. "Shure, an' Oi think she can. She's six feet hoigh, and four feet woide! Can she be seen? Sorrah a bit of anything ilse can ye see whin she's about."

Many a man's love for his club is due to the fact that his wife never gives her tongue a rest.

Figure 1: An example Trove news article showing the JSON representation overlaid with an image of the original scanned document, taken from `http://trove.nla.gov.au/ndp/del/article/64154501`

embedded in the newspapers for any particular location or time period.

The HuNI-Alveo project is a collaborative research project among researchers at Deakin University, University of Western Australia and Macquarie University. The aim of the project is to ingest Trove digitised newspapers into Alveo[2] virtual lab and to build Alveo's NER functionality to provide Trove-derived person or location names for ingestion into HuNI[3] virtual lab. To reach the second goal, we use the Stanford NER system (Finkel et al., 2005). A significant challenge in this project is to process the large number of news articles (approximately 152 million). We are not aware of any other work that applies NER to a collection of this size (the Europeana project (Willems and Atanassova, 2015) is of a similar size but there are no published NER results on the whole collection).

The remainder of this paper is organised as follows. In Section 2 we discuss our dataset and lexical resources that are used in the NER task. Section 3 represents evaluation results of the Stanford NER systems and Section 4 describes the NER pipeline that we implemented. Then, a series of interesting results are presented and analysed in Section 5. Section 6 describes how the results of the NER process are published as linked data on the web. Finally, conclusions and directions for future work are given in Section 7.

2 Data

The central ideas in the HuNI-Alveo project are to apply an NER model to historical newspapers to allow humanities researchers to exploit automatically identified person or location names. We use the following resources in this work.

2.1 Trove

Trove[4] is the digital document archive of the National Library of Australia (Holley, 2010) and contains a variety of document types such as books, journals and newspapers. The newspaper archive in Trove consists of scanned versions of each page as PDF documents along with a transcription generated by ABBYY FineReader[5], which is is a state-of-the-art commercial optical character recognition (OCR) system. OCR is inherently

[2] `http://alveo.edu.au/`
[3] `https://huni.net.au/`
[4] `http://trove.nla.gov.au/`
[5] `http://www.abbyy.com`

error-prone and the quality of the transcriptions varies a lot across the archive; in particular, the older samples are of poorer quality due to the degraded nature of the original documents. Generally errors consist of poorly recognised characters leading to mis-spelling or just random text in some cases. Article boundary detection seems to be very good.

To help improve the quality of the OCR transcriptions, Trove provides a web based interface to allow members of the public to correct the transcriptions. This crowdsourcing approach produces a large number of corrections to newspaper texts and the quality of the collection is constantly improving. As of this writing, the Trove website reports a total of 170 million corrections to newspaper texts[6].

As part of this project, a snapshot sample of the Trove newspaper archive will be ingested into the Alveo Virtual Laboratory (Cassidy et al., 2014) for use in language research. One motivation for this is to provide a *snapshot* archive of Trove that can be used in academic research; this collection won't change and so can be used to reproduce published results. Alveo also aims to provide access to the data in a way that facilitates automatic processing of the text rather than the document-by-document interface provided by the Trove web API.

The snapshot we were given of the current state of the collection contains around 152 million articles from 836 different newspaper titles dating from between 1803 and 1954. The collection takes up 195G compressed and was supplied as a file containing the document metadata encoded as JSON, one document per line. A sample document from the collection is shown in Figure 1 along with an image of the original page.

2.2 HuNI

The HuNI Virtual Laboratory (Humanities Networked Infrastructure `http://huni.net.au`) supports researchers in the Humanities to discover, document and link records about people, places and events in Australia. HuNI harvests data from many Australian cultural websites into a single data store.

One of the goals of this project was to provide HuNI with a new dataset linking names to articles in Trove. To facilitate this, HuNI provided an export of their current list of person records, around 288,000 records. Our goal was to find mentions of these people in Trove, rather than finding all names which we thought would be too large a data set. While each person record contains a list of attributes such as occupation and biography along with first-name/last-name pair, only a small fraction of records have both first name and last name. We built a name dictionary by extracting the names of persons who have both first and last names, leaving a total of 41,497 names.

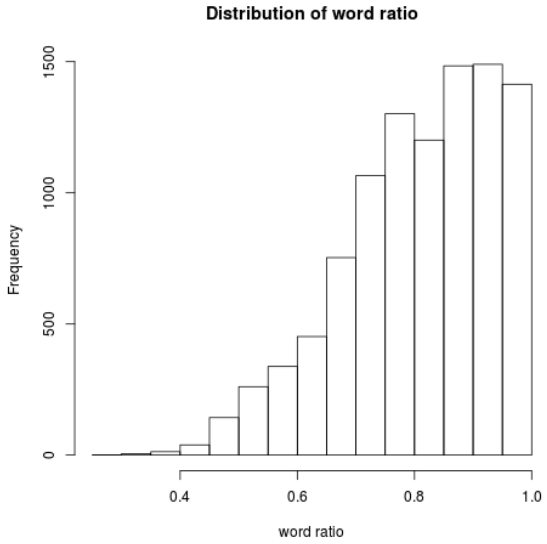


Figure 2: Histogram for the ratio of words to non-words over 10000 articles. The x-axis denotes the word frequency ratio and the y-axis denotes the number of articles.

2.3 Data Quality

As mentioned above, the quality of the OCR transcriptions in Trove is quite variable and we were concerned that the number of errors might be too high to allow useful results to be obtained from automatic processing. We thus investigate the quality of Trove in terms of word ratio with respect to a reference word list. The word list is derived from an Australian English dictionary [7] combined with the HuNI name list described above. Given an article, the word ratio is computed by dividing the number of words found in the dictionary by the total number of words in the article. This measures the relative frequency of words and non-words in

[6] `http://trove.nla.gov.au/system/stats?env=prod&redirectGroupingType=island#links`

[7] Derived from the Australian Learners Dictionary, available from `https://github.com/stevecassidy/ald`

the text . We assume that we can use word ratio as a proxy for OCR error rate and hence the quality of the text in each article (of course, many uncommon words will also be missing from the dictionary, making this measure an under-estimate of OCR error rate). Articles of poor quality would give a low word ratio, whereas articles of good quality would have high word ratio.

To evaluate the data, we randomly select 10000 articles[8] from the entire Trove dataset and estimated word frequency ratios over them. The histogram in Figure 2 shows the frequency ratio of words over 10000 articles. The x-axis denotes the frequency ratio of words and the y-axis denotes the number of new articles. We can observe the skew to the right in this small sample data which could indicate that the quality of the Trove data is not too bad. For instance, more than half the articles have a word frequency ratio greater than 0.8.

3 Evaluation

In this section we perform a comparative evaluation of two Stanford NER systems because we should make a decision about whether to train the NER system or not. To this end, we compare the performance of pre-trained Stanford NER with that of Stanford NER trained on our own training data. However, annotating data is a time-consuming and labour-intensive work and we thus use a semi-supervised learning approach. More specifically, training data is automatically generated using the pre-trained Stanford NER for randomly selected 600 articles and the produced silver standard data is used to train custom models for the Stanford NER system[9].

Some articles have a few sentences, even no names and they are not suitable for our evaluation. For this reason, we use the word frequency ratio described in Section 2.3 as a threshold to filter out inappropriate new articles. We randomly select 50 news articles from Trove that are not part of our training data using a word ratio threshold of 0.8. These articles were manually annotated using the MITRE annotation toolkit[10] to produce gold-standard test data for our evaluation.

On this test data, we evaluate the two Stanford NER systems and the comparison results are shown in Tables 1a and 1b. We can see that these two NER systems are on par with each other particularly in terms of F1 with respect to Person and Location, and our own trained Stanford NER does not provide any benefit. It would probably more desirable to use Stanford NER trained on more historical newspapers. However, this would be a labour-intensive and time-consuming task due to the huge amount of unannotated data. For these reasons, we use the pre-trained Stanford NER system, which gives us F1 scores of 0.76 for both person and location, in the rest of this paper.

As an aside, we also wondered if just using the HuNI supplied name list to look up names in the target articles would be a reasonable strategy. We ran an evaluation where words in target articles that were in the name list were tagged as PERSON instances. As might be expected with this approach, the recall is reasonable (0.75) since most of the target names will be found – errors are due to names not being present in the HuNI list. The precision though is very poor (0.07) since no cues are being used to differentiate ordinary words from names; hence, every occurrence of 'Carlton', 'rose' or 'brown' would count as a PERSON instance.

While we extracted and evaluated locations from the text, this paper concentrates on the use of person names. We hope to report on the application of location information in later work.

4 Extraction of Names

The goal of this work is to automatically extract person names and location names along with their relevant metadata from Trove. We use the pre-trained Stanford NER system that was evaluated in Section 3. The extraction of person names and their meta data is performed in four streaming steps as follows[11]:

1. Read news articles in Trove
2. Extract person entities from news context
3. Remove person names not found in the HuNI dictionary

[8] Actual number of articles is 9963 since 37 articles only have head information without article texts.

[9] We made a preliminary evaluation of Stanford NER given the increasing sizes of training data. We did not obtain any benefit from using more than 500 articles.

[10] http://mat-annotation.sourceforge.net/

[11] A noisy channel model was implemented to correct spelling errors in Trove but we did not obtain better quality texts using it. Furthermore, it seemed to be infeasible to apply it to the whole amount of Trove data due to extremely long processing time.

Entity	Precision	Recall	F1
Location	0.84	0.70	**0.76**
Organisation	0.56	0.47	0.51
Person	0.71	0.81	**0.76**
Totals	0.73	0.70	0.71

(a) Performance of pre-trained Stanford NER.

Entity	Precision	Recall	F1
Location	0.84	0.63	**0.72**
Organisation	0.54	0.28	0.37
Person	0.70	0.75	**0.73**
Totals	0.72	0.61	0.67

(b) Performance of Stanford NER trained on 600 articles.

Table 1: Performance comparison of Stanford NER systems in terms of precision, recall and f-score, figures quoted are micro-averaged.

4. Write tagged person named entities to a file

One of the most challenging issues in this work is to process large amounts of news articles, approximately 152 million articles as mentioned in Section 2.1. To tackle this issue, we implemented the extraction pipeline using a multiple threads to speed up processing. One extraction pipeline consists of several dedicated threads for reading, tagging and writing. In particular, multiple threads for tagging are used to communicate with multiple Stanford NER instances in a pipeline and this architecture leads to fast processing of large amounts of text. We utilised 15 virtual machines on the NeCTAR Research Cloud[12]; each machine was an m2.xlarge configuration with 48GB RAM and 12 virtual CPUs and the pipeline model ran on each virtual machine. The Trove data was divided into 30 chunks, each containing around 5 million news articles. Processing each chunk took 36 hours of processing time on average and the total processing time was about 72 hours.

The results contained 27 million person name mentions in 17 million articles; there were 731,673 different names - this includes some duplicates with different capitalisation.

Table 2 shows an example of the result for a name mention from Trove using the Stanford NER system; this includes some meta-data about the article containing the mention and a short text snippet showing the context of the first mention of the name in the article.

name: James Morgan, article_id: 13977910 article_date: 1894-11-30, article_source: The Sydney Morning Herald (NSW : 1842 - 1954), article_title: LEGISLATIVE ASSEMBLY. THURSDAY, NOVEMBER 29., article_context: ...n standing in tho name of Mr. James Morgan for tho appointment of a sole...,

Table 2: Extracted information for a person *James Morgan*.

5 Results and Analysis

This section shows some fundamental and interesting results and analysis[13] obtained from our NER system. The main aim of our project is to foster research on digital humanities through the use of NER and to deliver all necessary results for digital humanities scholars. The following sections describe several results that could be interesting and important topics for digital humanists working with historical texts.

5.1 Identifying Individuals

An important point to make here is that we are extracting *names* from the data, not *people*, however it is people that are of interest to Humanities researchers. Names are shared between many individuals over time as can be seen in Figure 3 which plots the occurence of the names of some Australian Prime Ministers for each year. Taking *Joseph Lyons* (red) as an example, there is a large peak in mentions around 1910 and a second peak in the 1930s. While these could refer to the same person, a little investigation shows that many of the 1910 mentions (eg. `http://trove.nla.gov.au/ndp/del/article/149796638`) refer to a Joseph Lyons arrested for destroying a railway line near Broken Hill (Figure 4). To make this data more useful to Humanities researchers it would be useful to be able to automatically cluster individuals within the data. This section describes one experiment in clustering names based on the in-

[12] `https://www.nectar.org.au/`

[13] Our results are publicly available via `http://trove.alveo.edu.au/` and we can perform a more detailed analysis using a query interface in SPARQL.

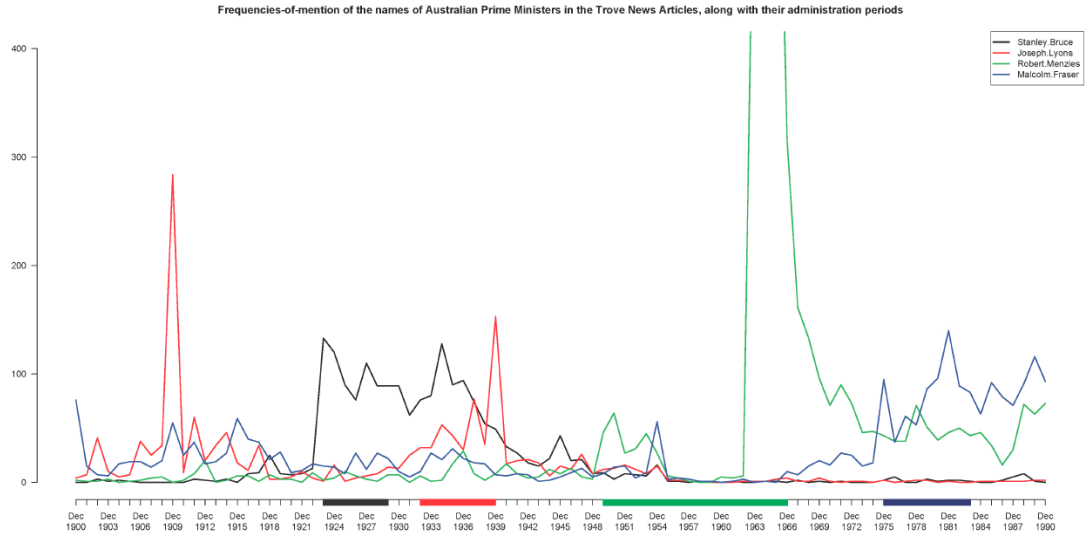
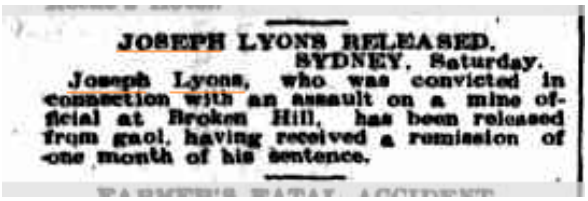


Figure 3: Frequencies-of-mention of the names of four Australian Prime Minsters along with their periods in office. Each colour zone on x-axis indicates the administration period of each Prime Minster.



JOSEPH LYONS RELEASED.
SYDNEY, Saturday.
Joseph Lyons, who was convicted in connection with an assault on a mine official at Broken Hill, has been released from gaol, having received a remission of one month of his sentence.

Figure 4: A mention of *Joseph Lyons* in 1909 which does not refer to the future Prime Minister (who was elected to the Tasmanian parliament in that year).

formation in the documents that mention them.

In this work we use a clustering approach on continuous vector space simply to distinguish whether the name *Joseph Lyons* belongs to the Australian Prime Minster or not. Previous work has proposed various approaches to represent words on the space such as latent semantic analysis (LSA) (Deerwester et al., 1990) or Latent Dirichlet Allocation (LDA) (Blei et al., 2003). In particular, the vector-space word representations learned by a neural network have been shown to successfully improve various NLP tasks (Collobert and Weston, 2008; Socher et al., 2013; Nguyen et al., 2015). Our work utilises the skip-gram model as implemented in freely available *word2vec*[14], which is a neural network toolkit introduced by Mikolov et al. (2013), to generate word vectors; they show that *word2vec* is competitive with other vector space models in capturing syntactic and semantic regularities in natural language when trained on the same data.

This work focuses on a name *Joseph Lyons* and we extract all news articles containing the name from Trove. For simplicity, we assume that there is only one *Joseph Lyons* for each year and the name is tagged with the publishing year of an article. For instance, *Joseph Lyons* of 1908 and *Joseph Lyons* of 1940 are represented as *joseph_lyons_1908* and *joseph_lyons_1940* in the extracted news articles, respectively. The total number of *Joseph Lyons* is 133 in this yearly representation. We train the *word2vec* skip-gram model on the extracted news articles and all the *Joseph Lyons* tagged with years are encoded to a 300-dimensional continuous word vector via the *word2vec* model.

The 300-dimensional word vectors of *Joseph Lyons* documents are projected into two-dimensional subspace using t-SNE (van der Maaten and Hinton, 2008) and clustered using the k-means clustering algorithm. We use the bayesian information criterion (BIC) to score the clusters for different values of k; the BIC score is maximum for $k = 4$ and so we select this number of clusters for *Joseph Lyons*. Finally we visualise the clusters on the plot based on the timeline as shown in Figure 5. The red line represents the period in office of Prime Minster *Joseph Lyons* and each colour zone on x-axis denotes one cluster in this figure. *Cluster4* is a close match to the true Prime Minister's time in office while *Cluster 3* shows another possible individual in the

[14] https://code.google.com/p/word2vec/

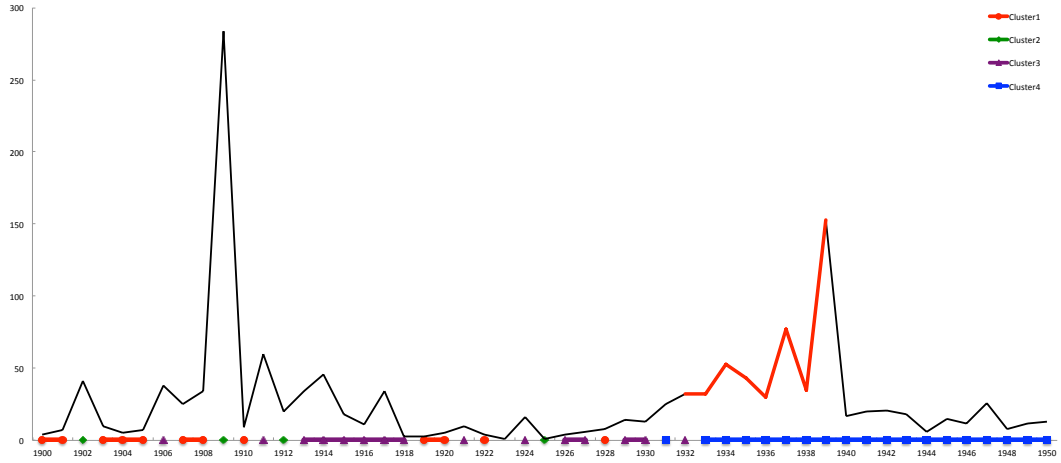


Figure 5: The frequency of mention of the name *Joseph Lyons* with cluster identifiers. The red line represents the period in office of Prime Minster *Joseph Lyons* and each colour zone on x-axis denotes one cluster.

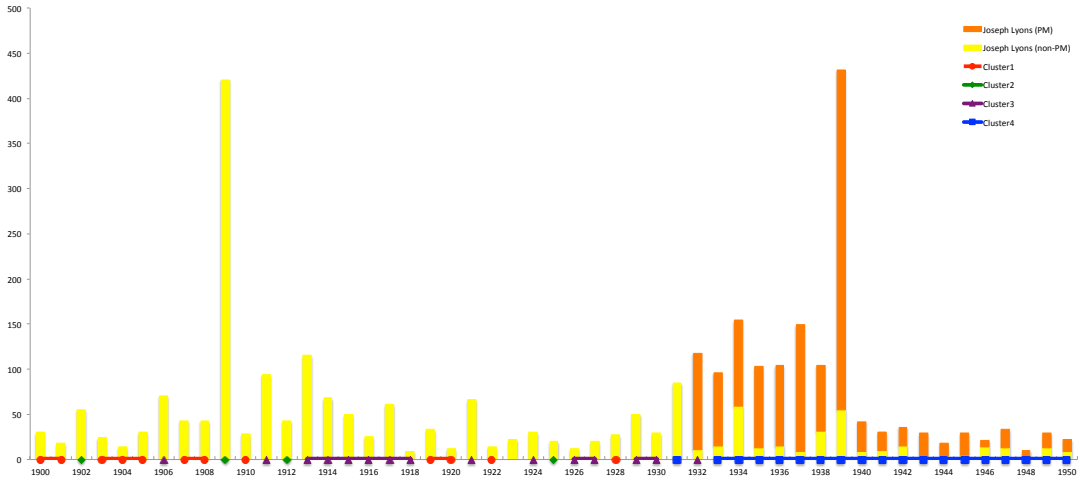


Figure 6: The number of news articles for each year mentioning *Joseph Lyons* as Prime Minster and non Prime Minster along with the clustering results from Figure 5.

period 1913-1918.

To validate the four clusters, we estimate the cluster purity by manually inspecting all news articles containing *Joseph Lyons* and counting those that refer to the PM and those that do not. Figure 6 plots the number of articles for each year mentioning *Joseph Lyons* as PM vs those that are not PM along with the identical clustering results shown in Figure 5. Note that we only count the number of articles of the Prime Minister *Joseph Lyons* and we do not take into account his previous political positions before becoming the Prime Minister.[15]

The figure shows that in the region of *Cluster4* the majority of mentions are of the PM while outside this region, the mentions are of a different individual (or *Joseph Lyons* before he was PM). Of the mentions in *Cluster4*, 75% are of the PM.

6 Publishing Linked Data

The results of the NER process have been made available to the HuNI project and will be integrated with their existing data collection as a new data feed linking names with Trove articles. However, we were interested in making a version of this data available in a way that would facilitate further experimentation and exploitation. To this end we have published a version of the data set on the web as *linked data*.

[15] *Joseph Lyons* successively held various Government positions before becoming the tenth Prime Minister of Australia. For instance, he became a new Treasurer of Australia in 1914.

```
<http://trove.nla.gov.au/ndp/del/article/60433109> a cc:Work ;
    dcterms:created "1919-01-10" ;
    dcterms:source <http://trove.alveo.edu.au/source/c987de65b64f0dab35715332478edccd> ;
    dcterms:title "Fatal Accident." ;
    schema:mentions <http://trove.alveo.edu.au/name/7e3030158f7e68d0e161feffd505ee60> ;
    trovenames:context "...hen Young, youngest son of Mr John Young, had the misfortune to meet w
    trovenames:year 1919 .

<http://trove.alveo.edu.au/name/7e3030158f7e68d0e161feffd505ee60> a trovenames:Name ;
    trovenames:word "john",
        "young" ;
    foaf:family_name "young" ;
    foaf:name "John Young" .
```

Figure 7: An example of a named entity mention converted to RDF in turtle format.

The principles of linked data (Berners-Lee et al., 2009) suggest that entities in the data set should be referenced by a URL and that this URL should resolve to a machine readable description of the entity that itself contains URL references to linked entities. A common underlying representation for linked data is RDF. To publish this data set we converted the named entity results to RDF using established vocabularies where possible. This version of the data is then hosted in an RDF triple store and a simple web application has been written to expose the data on the web.

An example of the RDF version of the data is shown in Figure 7. Trove articles are members of the class `cc:Work` and have properties describing publication date, title etc. Each article has one or more `schema:mentions` where each mention is an entity referring to a name. To facilitate searching, each name entity has properties containing the lowercase words in the name as well as the family name and the full name.

The resulting data set consists of 143 million triples and takes up around 26G of database storage using the 4store triple store[16]. A lightweight wrapper was written on this data to provide a web interface to the data set such that all of the URLs in the data resolve to the results of queries and return details of the particular resource. Using HTTP content negotiation, the response will be an HTML page for a web browser or a JSON representation for a script that sends an Accept header of `application/json`.

The API and the web application provide a SPARQL endpoint that supports queries over the data set. The web application is able to visualise the results of queries using the YASGUI[17] query front end.

As an example of mining the named entity data for more information, we wrote queries to find the *associates* of a given name. An associate is a name mentioned in the same document as another name. The query ranks the associated names by frequency of occurrence and returns the top 50 names. So, for example, the associates of *Robert Menzies* can be found at `http://trove.alveo.edu.au/associates/d857a2677bcb9955e286aafe53f61506` which shows that the top five are also politicians:

- Harold Holt (2552)
- Malcolm Fraser (1974)
- John Gorton (1596)
- Paul Hasluck (1232)
- John Curtin (1210)

This simple query shows some of the power that comes from having an accessible data source extracted from the Trove text. In the future we hope to be able to provide more kinds of query and visualisation that will enhance this data source for Humanities researchers.

7 Conclusion and Future Work

This paper has described a project to add value to a Humanities data set using standard NLP systems. The data set itself is interesting as a large collection of historical Australian newspaper text and will be made available via the Alveo virtual laboratory. Using a standard NER process we extracted 27 million person name mentions referencing 17 million articles in the archive. We have shown how this data can be exploited in a number of ways, namely by using a clustering method

[16] `http://4store.org/`
[17] `http://about.yasgui.org/`

to try to identify individuals in the data and by presenting the data set as linked data over the web.

The availability of this analysis has already proved interesting to Humanities researchers and we hope to be able to feed it back to the original Trove system run by the National Library of Australia. By providing this as an open data set the NLA encourage collaboration on the data and we hope to do the same with this new named entity data set.

References

Tim Berners-Lee, Christian Bizer, and Tom Heath. 2009. Linked data-the story so far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Julian Brooke, Adam Hammond, and Graeme Hirst. 2015. Gutentag: an nlp-driven tool for digital humanities research in the project gutenberg corpus. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 42–47, Denver, Colorado, USA, June. Association for Computational Linguistics.

Steve Cassidy, Dominique Estival, Tim Jones, Peter Sefton, Denis Burnham, Jared Burghold, et al. 2014. The Alveo Virtual Laboratory: A web based repository API. In *Proceedings of LREC 2014*, Reykjavik, Iceland.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pages 160–167, New York, NY, USA. ACM.

Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 363–370, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Rose Holley. 2010. Trove: Innovation in access to information in Australia. *Ariadne*, 64.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia, June. Association for Computational Linguistics.

Dat Quoc Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. 2015. Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics*, 3:299–313.

Olga Scrivner and Sandra Kübler. 2015. Tools for digital humanities: Enabling access to the old occitan romance of flamenca. In *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 1–11, Denver, Colorado, USA, June. Association for Computational Linguistics.

Richard Socher, John Bauer, Christopher D. Manning, and Ng Andrew Y. 2013. Parsing with compositional vector grammars. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 455–465, Sofia, Bulgaria, August. Association for Computational Linguistics.

Andrew J Torget, Rada Mihalcea, Jon Christensen, and Geoff McGhee. 2011. Mapping texts: Combining text-mining and geo-visualization to unlock the research potential of historical newspapers. *University of North Texas Digital Library*.

Laurens van der Maaten and Geoffrey E. Hinton. 2008. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605.

Marieke Willems and Rossitza Atanassova. 2015. Europeana newspapers: searching digitized historical newspapers from 23 european countries. *Insights*, 1(28):51–56.