Publishing the Trove Newspaper Corpus

Steve Cassidy
Department of Computing
Macquarie University
Sydney, Australia

Abstract

The Trove Newspaper Corpus is derived from the National Library of Australia's digital archive of newspaper text. The corpus is a snapshot of the NLA collection taken in 2015 to be made available for language research as part of the Alveo Virtual Laboratory and contains 143 million articles dating from 1806 to 2007.

This paper describes the work we have done to make this large corpus available as a research collection, facilitating access to individual documents and enabling large scale processing of the newspaper text in a cloud-based environment.

Keywords: newspaper, corpus, linked data

1. Introduction

This paper describes a new corpus of Australian historical newspaper text and the process by which it was prepared for publication as an online resource for language research. The corpus itself is of interest as it represents a large collection of newspaper text dating from 1806 to around 2007. However, to make this data available in a usable form, rather than a very large download, we have taken steps to build a usable web-based interface to the individual documents and to facilitate large scale processing of the text in a cloud-based environment.

2. The Trove Corpus

Trove[1] 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[2], which is is a state-of-the-art commercial optical character recognition (OCR) system. OCR is inherently 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.

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[3].

As part of this project, a snapshot sample of the Trove newspaper archive has been provided to 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 143 million articles from 836 different newspaper titles. The collection takes up 195G compressed and was supplied as a single file containing the document metadata encoded as JSON, one document per line. A sample document from the collection is shown in Figure 1.

2.1. Corpus Statistics

Figure 2 shows the number of documents for each year in the collection. While the overall range goes from 1806 to 2007, the majority of documents occur between 1830 and 1954; the latter date is due to the fact that most of the collection consists of out of copyright material with only a small amount of more recent material contributed by publishers. The range of counts goes from a few thousand per year to 3.5 million at the peak in 1915.

Figure 3 shows the average word length for documents for each year in the collection. There is some interesting variation in word length with earlier documents from the 1800s having almost double the word count on average to those in the 1900s. It is possible that this is due to inaccurate document segmentation from the scanned pages, but it perhaps reflects the change in stylistic norms for newspaper writing over time.

The overall number of documents in the corpus is around 144 million. In generating the statistics above we counted 69,568 million words based on the word count supplied with the source data. The tokenisation method that this count is based on is unknown and so a new tokenisation may derive a different number of words. The overall average document length is therefore 483 words.

2.2. Document Metadata

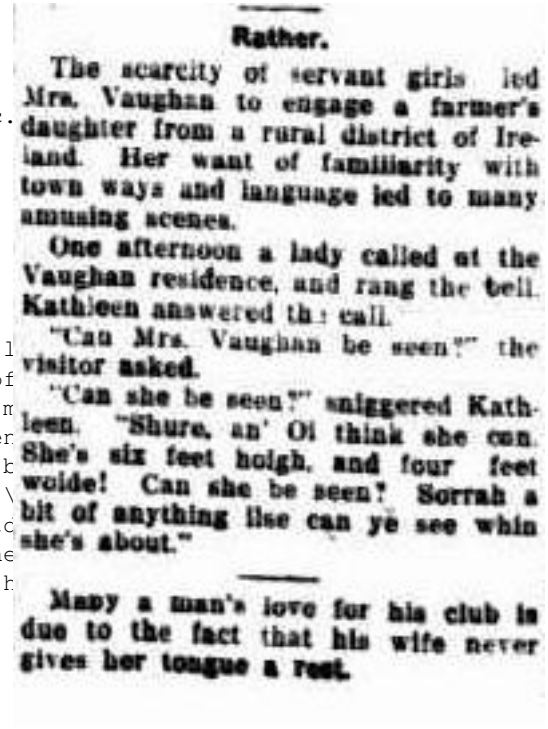
The corpus contains a wide range of document types from the wide range of newspapers included in the Trove archive. These range from small local newspapers to larger titles

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

[2] http://www.abbyy.com

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

```
{
 "id":"64154501",
 "titleId":"131",
 "titleName":"The Broadford Courier (Broadford, Vic.
 "date":"1917-02-02",
 "firstPageId":"6187953",
 "firstPageSeq":"4",
 "category":"Article",
 "state":["Victoria"],
 "has":[],
 "heading":"Rather.",
 "fulltext":"Rather. The scarcity of servant girls l
 engage a farmer's daughter from a rural district of
 of familiarity with town ways and language led to m
  One afternoon a lady called at the Vaughan resider
   Kathleen answered the call.' \"Can Mrs. Vaughan b
   asked. \"Can she be seen?\" sniggered Kathleen. \
   she can. She's six feet hoigh, and four feet Sotc
   Sorrah a bit of anything ilse can ye see whin she
   man's love for his club is due to the fact that h
   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

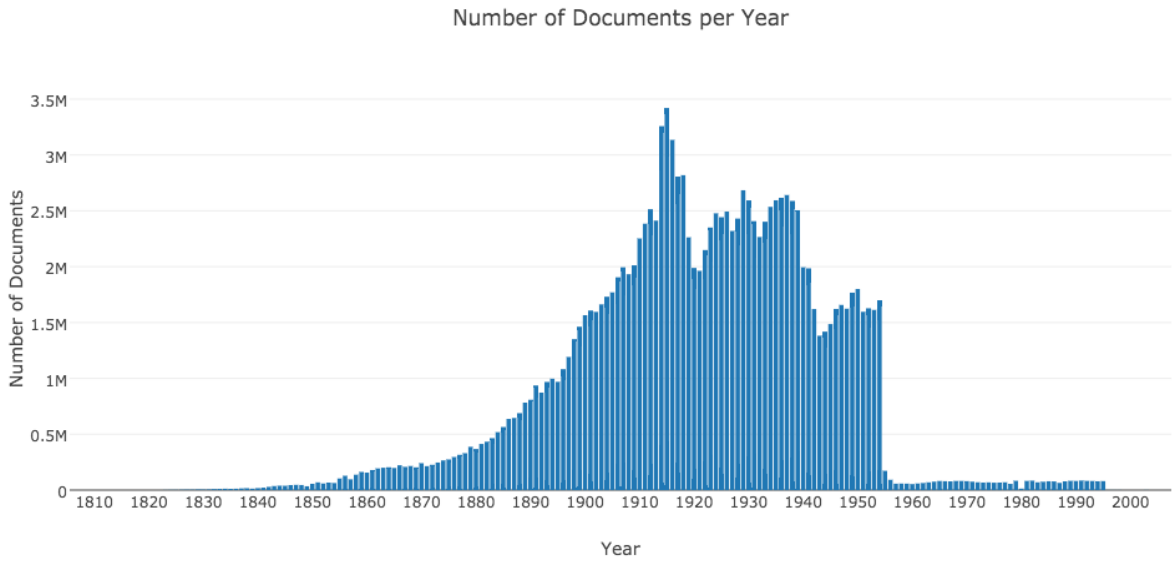


Figure 2: Number of documents per year in the Trove corpus.

from the major cities. A full list of the current holdings (which will be larger than the sample that we have) can be found on the NLA Trove website[4].

The document metadata includes the name of the newspaper and a *category* field with values such as *Article*, *Advertising*, *Detailed Lists, Results, Guides*, etc. Beyond this, there are no semantic categories associated with each document.

As mentioned earlier, NLA Trove website allows members of the public to correct the OCR generated transcriptions of newspaper articles. Our snapshot includes the most recent corrected version of the transcript and the metadata includes a flag to indicate that some corrections have taken place on the text. While it might be useful to have the original and corrected versions of the text, our current snapshot doesn't contain this information.

[4] http://trove.nla.gov.au/newspaper/about

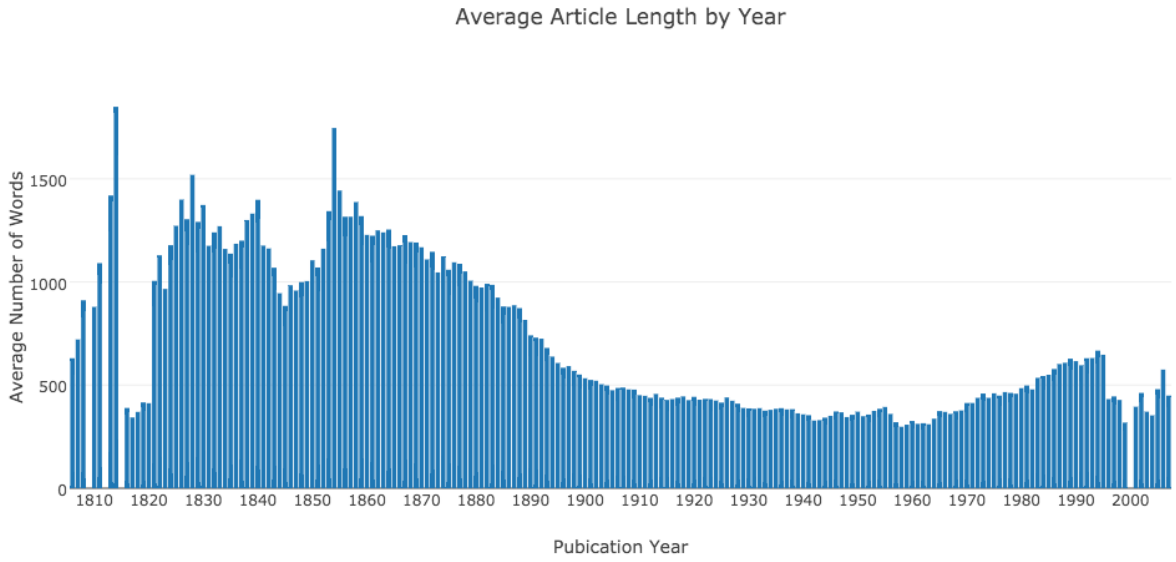


Figure 3: Average number of words per document for each year in the Trove corpus.

2.3. Related Corpora

Due to the importance of newspaper writing as a genre of written language, it has always been a part of text corpora used in language research. We can identify two broad categories of newspaper corpus: those that present a representative sample of newspaper writing and those that capture everything written over a specific period. The Trove collection falls into the latter category and includes all newspaper texts that have been contributed to the NLA project. Some examples of related corpora are listed here.

The Zurich English Newspaper corpus (Fries and Schneider, 2000) is a carefully curated representative corpus of early European newspaper text. Texts were sampled at roughly 10 year intervals from the earliest published newspapers in 1671 up to 1781. A total of around 1 million words was collected - analogous to the earlier Brown and LOB corpora.

The New Zealand Newspaper Corpus (Macalister and others, 2001) is another curated representative collection of newspaper texts sampled from the year 2000 and again containing around 1 million words. One of the goals of this corpus was to provide a representative sample of written NZ English in 2000 to form part of a larger collection of sampled texts starting in 1850.

The Reuters corpus (Rose et al., 2002) contains 806,791 articles from Reuters newswire and represents all English language stories published in the twelve month period starting in August 1996. This corpus is widely used as a reference collection in the information retrieval and natural language processing communities.

La Republica Corpus (Baroni et al., 2004) is a much larger collection of Italian newspaper text from 1995 to 2000 consisting of around 175 million words including all articles published in *La Republica* during that period. This corpus has been extended with a significant amount of annotation covering the structure of the text, part of speech, genre and topic tagging etc. Access to this corpus is provided via tools like the IMS Corpus Workbench and the Corpus Query Processor.

3. Publishing the Corpus

The goal of the project is to make the Trove Newspaper corpus available to language researchers in a way that makes it accessible for thier research.

The NLA Trove website allows full text search on the collection and shows the original scanned page image as well as the transcribed text. The Trove web API[5] provides a rich search interface and the ability to download a machine readable version of any document. Our goal in publishing the collection as a language resource is firstly to provide a stable snapshot as a basis for research, but also to provide interfaces that facilitate the kind of enquiry that language researchers need to carry out.

Publishing this collection is made difficult due to the large number of documents it contains. This section describes some of the issues that this has raised and our solutions to them.

3.1. Ingesting into Alveo

Alveo[6] (Cassidy et al., 2014) provides a data store for corpus data and a web based API that supports query over metadata and full text search. In many ways this is similar to the interface provided by Trove but Alveo is being used as the core of an eco-system for language processing tools; including the Trove collection in Alveo would make it available to researchers in the same way as all of the other

[5] http://help.nla.gov.au/trove/building-with-trove/api

[6] http://alveo.org.au

corpora. This means that, for example, search results displayed as concordances and word frequency counts are easily available from subsets of the Trove collection.

This is by far the largest collection that we have ingested into the system and it has exposed some inefficiencies in the ingest process of the Alveo system. Our previous work had concentrated on careful pre-processing of documents and generation of uniform meta-data from that supplied by the contributors. We estimated that ingesting Trove would take a number of months of processing time using the old system. To address this we have rewritten the ingest process of Alveo to use separate worker processes for metadata and full-text indexing driven by a message queue. This has significantly improved the ingest speed of the system to the extent that we can now ingest the collection in around five days.

3.2. Bulk Data Access

While the Alveo API provides some facilities for accessing larger collections of documents it does not yet provide a good interface for large scale data processing. Consequently, doing any computational work on the entire Trove corpus or even on significant subsets is not efficient using the Alveo (or Trove) web APIs. While we could just make a version of the entire collection available for download, we were interested in trying to develop an interface that would support this kind of processing using the research infrastructure available to us.

Australian researchers have access to an OpenStack based research cluster funded by the Government via NeCTAR [7]. OpenStack provides virtual machines and an object store through an interface similar to that of the Amazon EC2 and S3 services. This project leverages this infrastructure both to process and publish the Trove Newspaper collection but also to provide access to compute resources to allow researchers to make use of the data.

3.2.1. Data Format

The raw data was supplied to us by the NLA as a database dump with one JSON document per line where each JSON document has the same set of field names. This format is widely used in data processing and is often called the *dataframe* format. It is supported by tools such as R, Python (pandas) and large scale data processing tools such as Apache Spark[8].

3.2.2. Storing the Data

Alveo normally stores each document in a corpus as a file on disk and serves these through the API. With Trove, this would result in 143 million small files stored on disk which would require some work to optimise the file system in a default Linux volume. To solve this problem we split the original file into 150 uncompressed files of around 3G each and build an index containing the byte offset and length of each document. The uncompressed files are stored in the OpenStack object store (Swift, equivalent to Amazon S3); Swift provides HTTP access to these files including the ability to specify byte offsets and so retrieve the chunk of data corresponding to a single document or a number of contiguous documents. This enabled us to write a Python module[9] that provides access to individual documents and efficient iteration over large subsets of the corpus by retrieving chunks of data at a time.

Storing the data as a collection of 3G chunks means that algorithms that seek to process all of the data can be parallelised with each process working on one chunk of data. Using this method we are able to run simple processes on the data (e.g. word count) in a few hours on a single VM.

Since the data is stored in the line-by-line dataframe format, it is compatible with the facilities provided by Apache Spark which can then support parallel processing of the data within a chunk on a compute cluster. We are currently exploring this mode of access to the data to find a way to present a useable programming interface for end users.

Using the Python module we have written a web application[10] that makes individual documents available via an HTTP URL. To ingest data into Alveo, we use this URL to reference the document rather than feeding in the actual document; Alveo then redirects requests for the document to this web application rather than serving it from its internal file store.

These mechanisms provide a way of storing a very large textual data set, providing access to individual documents and a method for efficiently iterating over large groups of documents from a remote store.

4. Corpus Analysis

4.1. Data Quality

As mentioned above, the quality of the OCR transcriptions in Trove is variable, with many OCR errors evident when browsing through the data. To evaluate the quality of the data, we looked at the ratio of words appearing in a reference list to the total number of words per document. This metric only counts common words since our reference list is small, but gives some measure of the extent of OCR errors in the texts.

The word list is derived from an Australian English dictionary[11] containing about 10,000 words combined with a list of Australian first and last names with around 40,000 entries.

To evaluate the data, we randomly select 10000 articles[12] from the entire Trove dataset and estimated word frequency ratios over them. The histogram in Figure 4 shows the frequency ratio of words over these articles. We can observe

[7] http://nectar.org.au/research-cloud/

[8] https://databricks.com/blog/2015/02/02/an-introduction-to-json-support-in-spark-sql.html

[9] https://github.com/stevecassidy/trovenames

[10] http://trove.alveo.edu.au

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

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

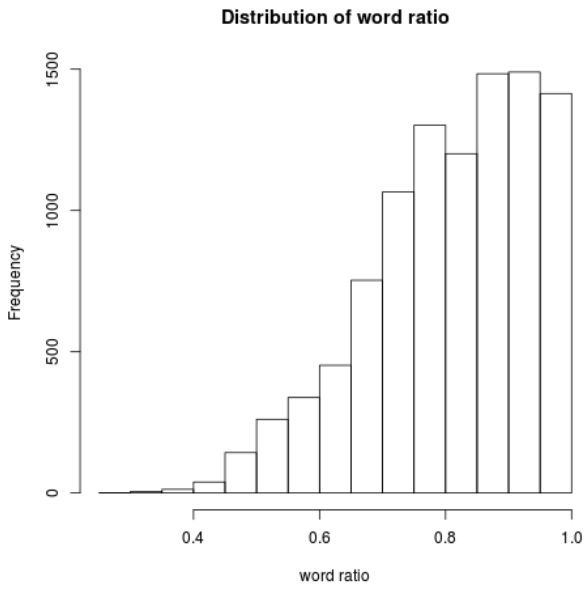


Figure 4: Histogram for the ratio of words to non-words over 10000 articles.

the skew to the right in this small sample which indicates that the quality of the Trove data is relatively good; more than half the articles have a word frequency ratio greater than 0.8.

4.2. Named Entity Recognition

As part of this project, we ran the Stanford Named Entity Recognition system over the Trove data to extract mentions of a large set of names (Mac Kim and Cassidy, 2015). The goal of this project was to link these names, which were of interest to Humanities researchers, to mentions of them in the Trove Newspaper archive.

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. Our evaluation of the accuracy of the NER process showed an F score of 0.76 on a small sample of hand labelled data.

To make the results of this process available, we converted the NER results to RDF and published them as linked data. The resulting data set consists of 143 million triples and takes up around 26G of database storage using the 4store triple store[13]. A lightweight wrapper was written on this data to provide a web interface to the data set that makes the results of the NER process available as Linked Data.

5. Future Work

The Trove corpus provides an unrivalled collection of newspaper text spanning a long period of time from the early days of publishing in Australia to the current century. As a historical resource, the NLA Trove interface is invaluable in supporting Humanities research. It is hoped that the publication of this snapshot as a linguistic resource will add a new dimension to this data in the research community.

[13] `http://4store.org/`

Our initial experiments with Named Entity Recognition show some useful results but also raise many interesting questions. For example, in finding target names of prominent Australians in the data we obviously see many instances of different individuals with the same name. The availability of such a large volume of data means that there are many instances of common names providing an excellent resource for looking at differentiating individuals based on the article context.

The wide range of dates represented in the corpus should provide a useful resource in looking at language variation over time; in particular since this represents a view over time of the evolution of Australian English from the early 1800s to the present day.

6. Summary

The Trove Newspaper corpus is a large collection of Australian newspaper text spanning the period from 1805 to around 1950. It is a snapshot of the digital archive of the National Library of Australia taken at a fixed point in time to provide a resource for language research. The size of the corpus presents a number of problems to making it available in a useful way to researchers. This paper describes the collection itself and the methods that we have used to make the data available to researchers via a web application.

The web application described in this paper is available at `http://trove.alveo.edu.au`.

7. Acknowledgements

This work was supported by a grant from Nectar, the Australian National eResearch Collaboration Tools and Resources project in collaboration with the Humanities Network Infrastructure (HUNI) project. The Trove archive is a fundamental piece of National research infrastructure provided by the National Library of Australia. The work was carried out in the context of the Alveo project, again funded by Nectar with development support from Intersect Australia. Work on Named Entity Recognition on the Trove corpus was carried out by Sunghwan Mac Kim.

8. Bibliographical References

Baroni, M., Bernardini, S., Comastri, F., Piccioni, L., Volpi, A., Aston, G., and Mazzoleni, M. (2004). Introducing the la repubblica corpus: A large, annotated, tei (xml)-compliant corpus of newspaper italian. *issues*, 2:5–163.

Cassidy, S., Estival, D., Jones, T., Sefton, P., Burnham, D., Burghold, J., et al. (2014). The Alveo Virtual Laboratory: A web based repository API. In *Proceedings of LREC 2014*, Reykjavik, Iceland.

Fries, U. and Schneider, P. (2000). Zen: preparing the zurich english newspaper corpus. *English Media Texts: Past and Present. Amsterdam: John Benjamins*, pages 1–24.

Holley, R. (2010). Trove: Innovation in access to information in Australia. *Ariadne*, 64.

Mac Kim, S. and Cassidy, S. (2015). Finding names in trove: Named entity recognition for australian historical newspapers. In *Australasian Language Technology Association Workshop 2015*, page 57.

Macalister, J. et al. (2001). Introducing a new zealand newspaper corpus. *New Zealand English Journal*, 15:35.

Rose, T., Stevenson, M., and Whitehead, M. (2002). The reuters corpus volume 1-from yesterday's news to tomorrow's language resources. In *LREC*, volume 2, pages 827–832.