Overview of the 2017 ALTA Shared Task: Correcting OCR Errors

Diego Mollá and **Steve Cassidy**
Department of Computing
Macquarie University
Sydney, Australia
diego.molla-aliod@mq.edu.au
steve.cassidy@mq.edu.au

Abstract

This paper presents an overview of the 8th ALTA shared task that ran in 2017. The task was to correct OCR errors from scans of newspapers stored in the Trove database maintained by the National Library of Australia. We introduce the task, describe the data and present the results of the participating teams.

1 Introduction

Many digital documents are the result of scanning printed copies. These documents, although in digital form, are in fact images, and as such, standard natural language processing techniques such as text search cannot be applied to them.

The National Library of Australia[1] maintains an archive of scanned Australian publications in the Trove database[2]. Many of these scans have been processed through Optical Character Recognition (OCR) and form a searchable resource with over 500 million items. But the OCR output may contain errors which need to be corrected. Trove has corrected the errors through a process of collaborative editing of the output of the OCR system.

The goal of the 2017 ALTA Shared Task is to automatically correct errors of OCR from a subset of scans from the Trove database. Over 7,000 documents were downloaded from the Trove database. For each document, the original output of the OCR system was used as the input text to the shared task, and the corrected versions were used as the target text. A total of 6,000 documents and their corrected versions were provided as the training set, and the rest was used to evaluate the system results.

This paper is structured as follows. Section 2 describes the shared task. Section 3 briefly introduces related research on OCR. Section 4 describes the data set that was used. Section 5 details the evaluation process. Section 6 presents and discusses the results. Finally, Section 7 concludes this paper.

2 The 2017 ALTA Shared Task

The 2017 ALTA Shared Task is the 8th of the shared tasks organised by the Australasian Language Technology Association (ALTA). Like the previous ALTA shared tasks, it is targeted at university students with programming experience. The general objective of these shared tasks is to introduce university students to the sort of problems that are the subject of active research in a field of natural language processing.

There are no limitations on the size of the teams or the means that they can use to solve the problem, as long as the processing is fully automatic — there should be no human intervention.

As in past ALTA shared tasks, there are two categories: a student category and an open category.

- All the members of teams from the **student category** must be university students. The teams cannot have members that are full-time employed or that have completed a PhD.
- Any other teams fall into the **open category**.

The prize is awarded to the team that performs best on the private test set — a subset of the evaluation data for which participant scores are only revealed at the end of the evaluation period (see Section 5).

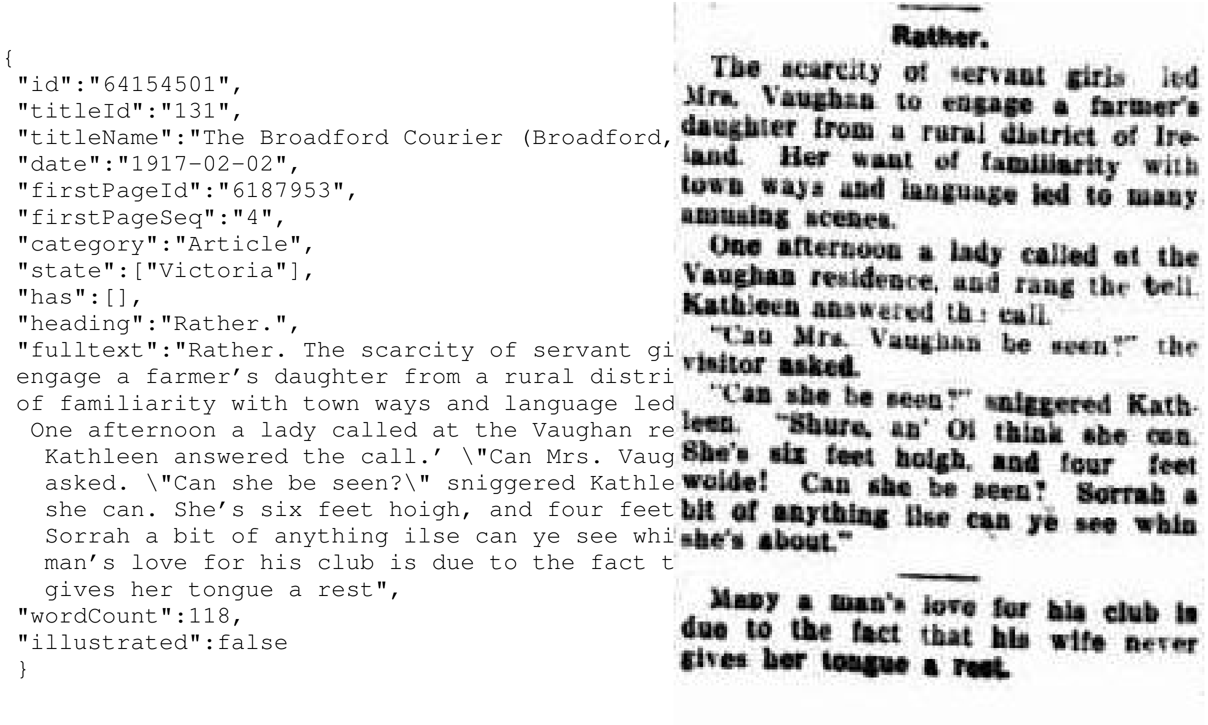
3 Related Work

OCR post-correction is a well established problem and has received some attention in particu-

[1] https://www.nla.gov.au/

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

Diego Mollá and Steve Cassidy. 2017. Overview of the 2017 ALTA Shared Task: Correcting OCR Errors. In *Proceedings of Australasian Language Technology Association Workshop*, pages 115−118.



```
{
 "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 th: 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

lar with reference to historical texts. Afli et al. (2016) describe two approaches to the problem. The first based on the application of a statistical machine translation system, treating the problem as one of translating the uncorrected OCR text into the corrected version. The second approach uses a language model to rank alternate corrections for words in the original OCR text. A *Noisy Channel Model* is used to model the errors introduced by the OCR process and the most probable correction is selected. Results on a corpus of ancient French manuscripts showed the best performance for the SMT based system with an word error rate of around 20%.

In contrast (Eger et al., 2016) make use of *character level* models of sequence mapping to correct OCR errors in Latin texts - interestingly they also apply the same methods to spelling correction in Tweets. This approach has the advantage of not requiring any kind of lexical model. The paper cites word error rates of the order of 10% for the Latin texts.

4 Data

Trove[3] 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[4], 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 crowd-sourcing 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[5].

The data for this evaluation was taken from a snapshot of the Trove Newspaper collection given to the Alveo Virtual Laboratory in 2015 (Cassidy, 2016) which consisted of 155 million individual documents. Some of these had already been cor-

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

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

[5] http://trove.nla.gov.au/system/stats

```
OBITUARY MR. J. G. KI.EMM i Mr .lohann Gottfried Klemm. Hi, of Gruenberg.
died on Saturday. He was l$>$orn on .lone $<>$, isa-j, being Tth child
...
```

Figure 2: Sample input provided by Trove.

```
<p><span> OBITUARY</span></p> <p><span> MR. J. G. KLEMM</span></p>
<p><span> Mr Johann Gottfried Klemm, 86, \&nbsp;</span><span>
of Gruenberg. died on Saturday.</span><span> He was born on
June 6, 1862, be-</span><span> ing 7th child ...
```

Figure 3: Sample target text provided by Trove.

```
OBITUARY MR. J. G. KLEMM Mr Johann Gottfried Klemm, 86, of Gruenberg.
died on Saturday. He was born on June 6, 1862, being 7th child
...
```

Figure 4: Target text after it has been cleaned automatically.

rected and this was recorded in the metadata for each document. For this evaluation, we selected a subset of documents that had no corrections in the 2015 snapshot and for each of these used the Trove API to retrieve the most recent (July 2017) version of the document. Where this newer version contained some manual corrections we included the document pair in the collection.

Figures 2 and 3 show a sample input text and the target text, respectively, as they were provided by Trove. Note the presence of XML markup and the occurrence of words that were split across two spans in the target text. A Python script was used to clean the target text, giving the result of Figure 4 for the example of Figure 3.

Given the nature of the process used to produce the annotations, some errors remained in the final annotations. In particular, not all of the OCR errors of the input text had been corrected by the annotators. In addition, in a number of cases the text provided by Trove included words spanning two lines which were not hyphenated. These words would appear as two separate (incorrect) words in the target text.

The training data contained 6,000 documents. The test data contained 1,941 documents.

5 Evaluation

As in previous ALTA shared tasks, the 2017 shared task was managed and evaluated using Kaggle in Class, with the name "ALTA 2017 Challenge". The Kaggle in Class framework allowed the maintenance of a discussion forum that could be used to communicate among the participants. In addition, thanks to this framework the participants were able to submit runs prior to the submission deadline for immediate feedback.

The test data was partitioned into a public and a private section. Whenever a participating team submitted a run, the evaluation results of the public partition were immediately available to the team, and the best results of each team appeared in a public leaderboard. The evaluation results of the private partition were available to the competition organisers only, and were used for the final ranking after the submission deadline. To split the test data into the public and private partitions, we used the defaults provided by Kaggle in Class. These defaults performed a random partition with 50% of the data falling into the public partition, and the remaining 50% falling into the private partition. The participants were able to see the entire unlabelled evaluation data, but they did knot know what part of the evaluation data belonged to which partition.

Each participating team was allowed to submit up to two (2) runs per day. By limiting the number of runs per day, and by not disclosing the results of the private partition, the risks of overfitting to the private test results were diminished.

The chosen evaluation metric was the mean F1 score. This metric is common in information retrieval tasks, and measures the harmonic mean of recall and precision according to the formula:

$$F1 = 2\frac{p \cdot r}{p + r}$$

Where p is the precision and r is the recall. Recall and precision were computed at the level of bigrams. By operating on bigrams instead of single words, the metric was sensitive to differences of word order.

Furthermore, the participants were asked to remove all bigram duplicates and all bigrams from the solution already occurring in the original text prior to submission. The participants were provided with a Python script that removed such information. By removing all bigrams already occurring in the original text, the evaluation focused on words that were corrected by the systems. This was important, since otherwise a trivial system that did not perform any OCR correction and simply returned the input text unmodified would have achieved an F1 score of 84.6% because many words of the input text do not require correction.

A further constraint on the output was that each word forming a bigram should not contain quotation marks or blank spaces. This constraint was required due to the CSV format used by the files that were processed by the evaluation scripts from Kaggle in Class. The Python script provided to the participants also removed these problematic bigrams.

6 Results

Table 1 shows the results of the public and private partitions for all participating teams. The results in the public and private partitions were consistent, and team EOF was a clear winner.

Table 1: F1 of all participating systems.

System	Category	Public	Private
EOF	Student	0.33497	0.32987
SuperOCR	Student	0.16798	0.16817
Atom	Student	0.14127	0.14654
CTexT	Open	0.08539	0.08625
Natural Language	Student	0.02768	0.02610

7 Conclusions

The 2017 ALTA Shared Task was the 8th of the series of shared tasks organised by ALTA. This year's shared task focused on OCR correction, and the data was extracted from the Trove database maintained by the National Library of Australia.

The crowdsourcing nature of the annotation process, and the format returned by Trove, caused a number of annotation errors which make this task particularly challenging to the participating teams.

For full details of some of the participating systems, refer to the shared task section of the 2017 ALTA workshop proceedings.

References

H Afli, Z Qiu, A Way, P Sheridan LREC, and 2016. 2016. Using SMT for OCR error correction of historical texts. *computing.dcu.ie* .

Stephen Cassidy. 2016. Publishing the trove newspaper corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. European Language Resources Association (ELRA).

Steffen Eger, Tim Vor der Brück, and Alexander Mehler. 2016. A comparison of four Character-Level String-to-String translation models for (OCR) spelling error correction. *The Prague Bulletin of Mathematical Linguistics* 105(1):781.

R Holley. 2010. Trove: Innovation in access to information in australia. ariadne, 64. *www.ariadne.ac.uk/issue64/holley* .