Named Entity Recognition for Digitised Historical Texts

Claire Grover[*], **Sharon Givon**[*], **Richard Tobin**[*] **and Julian Ball**[†]

[*]School of Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW, UK
grover@inf.ed.ac.uk, S.Givon@sms.ed.ac.uk,
richard@inf.ed.ac.uk

[†]BOPCRIS
University of Southampton
Hartley Library
University Road
Southampton SO17 1BJ, UK
J.H.Ball@soton.ac.uk

Abstract

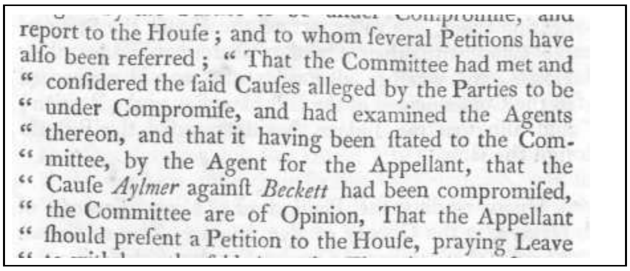
We describe and evaluate a prototype system for recognising person and place names in digitised records of British parliamentary proceedings from the late 17th and early 19th centuries. The output of an OCR engine is the input for our system and we describe certain issues and errors in this data and discuss the methods we have used to overcome the problems. We describe our rule-based named entity recognition system for person and place names which is implemented using the LT-XML2 and LT-TTT2 text processing tools. We discuss the annotation of a development and testing corpus and provide results of an evaluation of our system on the test corpus.

1. Introduction

In this paper we describe how we have applied text processing techniques to historical texts from the BOPCRIS 18th Century Parliamentary Publications project (`http://www.bopcris.ac.uk/18c/`) which makes digitised parliamentary records widely accessible via the web. Pages from the records are first scanned and then converted to text using optical character recognition (OCR) technology so that they can be indexed and searched. The results of search are displayed back to the user in the form of highlighting of terms in the jpeg images of the pages. The premise behind the research reported here is that more sophisticated indexing and search may be possible if particular types of terms can be automatically identified in the documents. Our role has been to apply text processing and named entity recognition (NER) techniques for the automatic recognition of person and place names. We have experience of building both rule-based NER systems (Mikheev et al., 1999a; Mikheev et al., 1999b) and ones which are machine-learning-based (Alex et al., 2006; Finkel et al., 2005; Hachey et al., 2005; Nissim et al., 2004). For the initial prototype described here we decided to use rule-based methods. In part this decision was taken to avoid the cost of annotation of training data, but it was also taken because it was not obvious that machine-learning methods would lend themselves to this data. Developing a rule-based system has allowed us to explore what is rather unusual and problematic data, as discussed in the next section.

ßands (stands)	*feve-rälly* (seve-rally)
Erßine (Erskine)	*haYe* (have)
$Ap\neg\ \ peals$ (Appeals)	*refpecT:* (respect)
$Lord\neg\ \ fhips$ (Lordships)	*0 ’ Done 1* (O’Donel)

Figure 1: OCR errors in the 1814–1817 data set

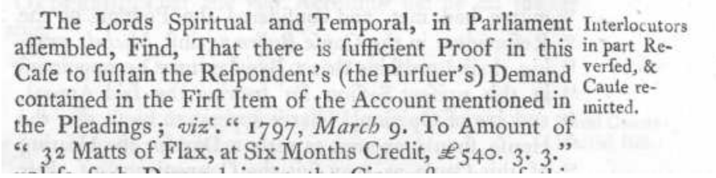


report to the Houſe ; and to whom ſeveral Petitions have
alſo been referred ; “ That the Committee had met and
“ conſidered the ſaid Cauſes alleged by the Parties to be
“ under Compromiſe, and had examined the Agents
“ thereon, and that it having been ſtated to the Com-
“ mittee, by the Agent for the Appellant, that the
“ Cauſe *Aylmer* againſt *Beckett* had been compromiſed,
“ the Committee are of Opinion, That the Appellant
“ ſhould preſent a Petition to the Houſe, praying Leave

Figure 2: Use of quotation marks

2. The Parliamentary records data

We have used two sets of data from the Journals of the House of Lords, an early set from 1685 to 1691 and a later set from 1814 to 1817. The two sets were OCRed at different times using different processors and, in fact, the quality of the OCR of the earlier documents is better because the process was specialised to Old English and it deals properly with the ‘f’-like realisation of ‘s’ that occurs in both sets. In general, OCR quality is erratic and there is a tendency for unusual non-alphabetic characters to appear, as illustrated in Figure 1. Problems also arise with the use of quotation marks. The texts are densely populated with quotations and every line of a quotation is preceded by a double quote mark, as in Figure 2. These quote marks frequently interfere with the natural flow of words in a named entity and special steps are needed to deal with them. Moreover, the OCR has problems identifying these double quotes and outputs them as a wide range of non-standard characters. In addition, the OCR system has problems with layout and is frequently unable to distinguish marginal notes from the main body of the text, giving rise to discontinuous sequences of words which could confound a NER system—Figure 3 shows an example of this problem. Although these are English texts, there are frequent portions in Latin, especially in the 1685-1691 data, and OCR seems to be poorer on these sections than on the English sections.



The Lords Spiritual and Temporal, in Parliament
aſſembled, Find, That there is ſufficient Proof in this
Caſe to ſuſtain the Reſpondent's (the Purſuer's) Demand
contained in the Firſt Item of the Account mentioned in
the Pleadings; *viz*^{t}. " 1797, *March* 9. To Amount of
" 32 Matts of Flax, at Six Months Credit, £540. 3. 3."

Interlocutors
in part Re-
verſed, &
Cauſe re-
mitted.

```
   The Lords Spiritual and Temporal, in Parliament Interlocutors affembled, Find,
That there is fufficient Proof in this in Part Re-Cafe to fuftain the Refpondent's
(the Purfuer's) Demand Jjjjj⌒contained in the Firft Item of the Account mentioned
in mitted.  the Pleadings; viz*." 1797, March 9.  To Amount of " 32 Matts of Flax,
at Six Months Credit,⌒540.  3.3."
```

Figure 3: OCR problems with marginal notes

Aside from OCR issues, the Parliamentary data is difficult in other respects and it is unclear whether the character-based features which work well in machine-learning NER systems would carry over well to this data. A major concern is the different use of word-initial capitals where common nouns are more likely to be capitalised than not (see for example, Figures 2 and 3). The presence or absence of capitalisation is a highly significant feature for machine-learning NER systems and a German-style initial capitalisation of nouns can significantly impact on performance. In CoNLL 2003 (Sang and De Meulder (2003)) NER systems were trained and tested on data from English and German. The highest f-score achieved for German was 72.41 as compared to 88.76 for English and the different capitalisation is likely to be the main reason for this difference. A further issue concerns part-of-speech (POS) tagging. The peculiar nature of our data means that a modern English POS tagger is unlikely to perform well on it, and this in turn would have an impact on the quality of a machine-learning NER system. These considerations meant that we had concerns about whether a machine-learning approach would achieve a high-level of accuracy on the BOPCRIS data. For this reason, and because of the costs of annotation of training material, we decided to pursue rule-based techniques.

3. Data annotation

Although a rule-based approach does not require annotated training data, gold standard annotated material is necessary for development and testing. We therefore annotated samples of data from the 1814–1817 and the 1685–1691 data sets, using guidelines developed specifically for the task. In total 91 files from the 1814–1817 set were annotated. There are 136,646 words in these 91 files and 2,751 person names and 2,021 location names were annotated. The data was split into two sets, devtest and test (45 and 46 documents respectively). The devtest set was created for system development and could therefore be seen by developers. The test set was created for formal evaluation. The 1685–1691 data set was processed with Old English OCR and we annotated 46 documents from this collection to make a second test set. This enabled us to discover whether there would be a difference in performance on data that comes from an earlier period and that was OCRed in a slightly different way. There are 51,901 words in these 46 files and 3,199 person names and 164 location names were annotated. Place names occur significantly less frequently in this set.

	1814–1817 Set	1685–1691 Set	Both
Location	85.91	65.03	80.75
Person	93.57	95.10	94.03
Total	90.66	93.06	91.46

Table 1: Inter-annotator agreement (f-score)

We used the WordFreak annotation tool (`http://wordfreak.sourceforge.net/`) and marked up three entity types, *person*, *location* and *interrupt*. The *interrupt* entity was used to mark material that occurred inside a person or place name but which was not part of it, for example, marginal notes or quotation marks as discussed in the previous section. We doubly annotated 16 randomly chosen files from the test sets of each of the two data sets and calculated inter-annotator agreement (IAA) using balanced f-score. Table 1 shows the results for both sets, separately and combined.

In general, IAA is high but it is significantly higher for person names than for location names. The lowest score is for location names in the 1685–1691 set: this may in part be due to data sparseness in this set, as there were only 164 location names annotated in the 46 test files as opposed to 2,021 location names in the corresponding 1814–1817 test set. A further problem with the 1685–1691 set was that one annotator systematically annotated names within the marginal notes while the other did not. This was discovered after IAA was calculated and was subsequently rectified. The test sets used for evaluation contain arbitrated versions of the doubly annotated files.

4. The prototype system

The NER tagger system developed for this project was implemented using in-house XML tools. Over the past few years we have developed a suite of tools for generic XML manipulation (LTXML, Thompson et al. (1997)) as well as NLP specific XML tools (LT TTT, Grover et al. (2000)). More recently we have developed significantly improved upgrades of these tools, LT-XML2 and LT-TTT2 (`http://www.ltg.ed.ac.uk/software/ltxml2`).

The input to the system is an XML format output by the OCR process. The first step in the processing pipeline converts this to a more suitable format and performs transformations such as conversion to utf-8, separation of trailing punctuation and white space from word elements and addition of ids to word elements. Each word element in the original has coordinate information associated with it in the attributes l (left) and t (top) and size information in the attributes h (height) and w (width). This information is preserved and is used to calculate new information such as the location of newline white space. We also attempt to deal with the problem of marginal notes: as described above, the material in the right-hand margin is frequently intermingled with the preceding column. We use the coordinate information to try to identify w elements which are located in the margin and add to them the attribute `marginal='true'`. This attribute is later used by the named entity recognition component to filter marginal words out of names.

```
<entities>
 <person>
  <name>Mr. Stratford</name>
  <regions>
   <region h="30" l="417" t="881" w="59"/>
   <region h="41" l="489" t="881" w="146"/>
  </regions>
 </person>
 <location>
  <name>County of Kent</name>
  <regions>
   <region h="40" l="751" t="1878" w="119"/>
   <region h="31" l="885" t="1875" w="35"/>
   <region h="32" l="931" t="1874" w="117"/>
  </regions>
 </location>
</entities>
```

Figure 4: System output

The NER tagger is implemented as a sequence of calls to the LTG's LT-XML2 programs, in particular the XML transduction program *lxtransduce*. This program operates on an XML input file to add person and place name markup as defined by rules in grammar files. For person names there are rules for, amongst others, monarchs, earls, lords, dukes, churchmen, and common people. The rules access a variety of lexicons including ones listing male and female christian names and surnames, as well as more specialist lexicons such as one which lists place names which are also earldoms (e.g. “Warwick”).

Place name recognition is done near the end of the NER tagging pipeline and is slightly interleaved with person name recognition. After all the specialist rules for dukes etc. have applied, a first set of place name rules are used to identify very high confidence place names, such as “Town of London”. Then a general purpose person name grammar operates to find common names such as “Mr. Stratford” before a general purpose place name grammar finishes the process. Because of this incremental approach to adding mark-up, the specialist lexicons, such as the one for place names which are earldoms, are used first to ensure that an example such as “Earl of Warwick” is recognised as a person name.

Before the entity rules are applied, a grammar to identify possible ‘noise’ is used. This marks marginal notes, quotation marks, unusual characters etc. as noise so that the entity grammar rules can permit noise to occur in the middle of an entity. At the end of the NER tagging pipeline the noise mark-up is removed since it has now served its purpose.

The NER tagger outputs a file which is the same as the pre-processed input file except that additional entity markup has been added. This is then converted to a ‘standoff’ file containing a list of all the entities found in the page, as shown in the small example in Figure 4. Here the *regions* element lists regions which correspond to the word elements from the original input file and which have inherited their h, l, t and w attributes. This output format is designed to enable highlighting of entities in the page image through a process of mapping from the regions in the OCR XML format to regions in the image.

5. Evaluation

We have evaluated the system both on the devtest set (from the 1814–1817 data) as well as on the blind test sets from the two time periods. The results are shown in Table 2 where precision (P), recall (R) and f-score (F) are shown.

The scores on the 1814–1817 sets are reasonable for a first prototype produced within a limited time span, especially when we consider the historical nature of the data and the fact that it comes from a noisy OCR source. Recognition performs better for person names than for location names and this may reflect the fact that person name recognition is more dependent on finding patterns in the text while location name recognition is more dependent on gazetteer resources: the former is more resistant to OCR error than the latter. To illustrate, consider the string ‘Earl of Shagefiury’ (Earl of Shaftesbury). The rules can identify this as a person name, even though the OCR has mangled ‘Shaftesbury’, because the ‘Earl of capitalised-word’ pattern can still match. If the string ‘Shagefiury’ appeared on its own as a place name, however, a look-up against the gazetteer would fail and the name would be missed. Solutions to this problem lie either in better OCR technology or in fuzzy look-up techniques.

The results for the 1685–1691 test set are comparable to the 1814–1817 data for person names but are significantly worse for location names. Note that while precision and recall have both fallen considerably for locations, the largest drop is in precision. This indicates that the rules are frequently predicting location names where they were not annotated in the gold standard. In general, the low incidence of location names in this set indicates that it is quite different from the 1814–1817 set that the location name rules were developed for. Moreover, IAA was lower for locations in this set, indicating that the location decision is in some way harder than it was for the first set. There are other contributing factors, including differences in naming conventions between the two sets. For example, the title ‘Vicecomes’ does not occur in the devtest set and was not

	1814–1817 DevTest			1814–1817 Test			1685–1691 Test		
	P	R	F	P	R	F	P	R	F
Location	71.94	72.29	72.12	67.92	65.19	66.53	16.59	43.90	24.08
Person	81.81	79.56	80.67	79.07	72.42	75.60	81.65	69.40	75.03
Total	77.47	76.42	76.94	74.34	69.44	71.81	72.69	68.15	70.35

Table 2: NER Tagger Evaluation Results (%)

included in the rules or lexicons but it appears frequently in the 1685–1691 set. Thus 'Vicecomes Weymouth' is not recognised as a person name and 'Weymouth' is then incorrectly identified as a location. OCR interpretation of layout also seems to be a contributing factor. For example, a column which is a list of bishop's names of the form 'Epus. Durham.' is wrongly recognised as two columns with 'Epus.' and 'Durham.' separated. This means that the relevant person name rule doesn't fire and the instance of 'Durham' is recognised incorrectly as a location. (This problem does not affect the recall of person name recognition as the annotators did not attempt to mark-up names when the columns were severely mangled.)

Further error analysis of the 1814-1817 devtest data reveals a number of problems, many arising from the OCR process. One source of problems is words broken into more than one w element, e.g., "Sonderland" was broken into `<w>Sonde</w> <w>rland</w>`. Some of the instances of interrupting noise can be handled by the grammar rules, but where lexical look-up is needed it can fail because of noise or because of misrecognition of characters. For example, "Lancqßire" is split into three w elements, "Lancq ß ire" and additionally the 'a' is misrecognised as a 'q' so that matching against the lexical entry "Lancashire" fails. The OCR process also sometimes creates bad paragraph breaks in the middle of names. which prevents them from being detected. For example, there is a paragraph split after "Neal" in the name "Sir Neal O'Donel". Note that the OCR output also contains sentence mark-up but this is so unreliable that we discard it.

6. Conclusions and future directions

The evaluation results are promising and we intend to improve performance by continuing to develop the rules and lexicons. In addition to continued development, there are other questions and extensions that it would be useful to consider. The main problem is clearly the low quality of the OCR output and it would be fruitful to explore methods of separating the main text and the marginal notes more clearly. In addition the recognition of individual characters is very noisy and it might be interesting to explore ways of correcting this. One route would be to pursue machine-learning methods to perform automatic correction of the OCR output. This would require a collection of parallel data which had been both OCRed and retyped to use as training material and would utilise automatic alignment methods of the kind used in machine translation systems. Due to limited time and resources we confined our attention to person and place names. However, it would be interesting to extend the system to other entities such as organisations and dates. Further extensions to relation and event extraction could also be implemented.

7. Acknowledgements

This research was carried out as part of the BOPCRIS PARL18C project, funded by the UK Joint Information Systems Committee (JISC). We would also like to thank Laurence Durnan, Ed Fay, Barry Haddow, Ian Hughson and Maria Milosavljevic for their assistance with this project. The annotated data described here is available from `http://www.ltg.ed.ac.uk/` for research use only.

8. References

Beatrice Alex, Malvina Nissim, and Claire Grover. 2006. The impact of annotation on the performance of protein tagging in biomedical text. In *Proceedings of the Fifth Language Resources and Evaluation Conference (LREC2006)*.

Jenny Finkel, Shipra Dingare, Christopher D. Manning, Malvina Nissim, Beatrice Alex, and Claire Grover. 2005. Exploring the boundaries: Gene and protein identification in biomedical text. *BMC Bioinformatics*, 6(Suppl1):S1.

Claire Grover, Colin Matheson, Andrei Mikheev, and Marc Moens. 2000. LT TTT—a flexible tokenisation tool. In *LREC 2000—Proceedings of the 2nd International Conference on Language Resources and Evaluation*, pages 1147–1154.

Ben Hachey, Beatrice Alex, and Markus Becker. 2005. Investigating the effects of selective sampling on the annotation task. In *Proceedings of the 9th Conference on Computational Natural Language Learning*.

Andrei Mikheev, Claire Grover, and Marc Moens. 1999a. XML tools and architecture for named entity recognition. *Journal of Markup Languages: Theory and Practice*, 1(3):89–113.

Andrei Mikheev, Marc Moens, and Claire Grover. 1999b. Named entity recognition without gazetteers. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL'99)*, pages 1–8.

Malvina Nissim, Colin Matheson, and James Reid. 2004. Recognising geographical entities in Scottish historical documents. In *Proceedings of the Workshop on Geographic Information Retrieval, SIGIR ACM 2004*.

Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the 2003 Conference on Computational Natural Language Learning*.

Henry Thompson, Richard Tobin, David McKelvie, and Chris Brew. 1997. LT XML. software API and toolkit for XML processing. *`http://www.ltg.ed.ac.uk/software/`*.