"Any more Bids?": Automatic Processing and Segmentation of Auction Catalogs

West, Kris
Kris.West@gmail.com
JSTOR, USA

Llewellyn, Clare
Clare.Llewellyn@ithaka.org
JSTOR, USA

Burns, John
John.Burns@ithaka.org
JSTOR, USA

This paper details work that has been conducted through a collaborative project between JSTOR, the Frick Collection and the Metropolitan Museum of Art. This work, funded by the Andrew W. Mellon Foundation, was to understand how auction catalogs can be best preserved for the long term and made most easily accessible for scholarly use. Auction catalogs are vital for provenance research as well as for the study of art markets and the history of collecting. Initially a set of 1604 auction catalogs, over 100,000 catalog pages, was digitised – these catalogs date from the 18th through the early 20th century.

An auction catalog is a structured set of records describing items or lots offered for sale at an auction. The lots are grouped into sections – such as works by a particular artist, each of the sections are then grouped into a particular sale – this is the actual event that happened in the sale room, and then these sales are grouped together in the auction catalog. The auction catalog document also generally includes handwritten marginalia added to record other details about the actual transaction such as the sale price and the buyer.

A repository was constructed – this holds and provides access to page images, optical character recognition (OCR) text and database records from the digitised auction catalogs. In addition a website was created that provides public access to the catalogs and automatically generated links to other collections. This site offers the ability to search and browse the collection and allows users to add their own content to the site.

When searching a user may only be interested in a single item within a larger catalog, therefore, to facilitate searching the logical structure of the catalog needs to be determined in order to segment the catalog into items. The catalogs are extremely variable in structure, format and language, and there are no standard rules that can divide the catalog into the lots, sections and sales. Therefore, machine-learning techniques are used to generate the segmentation rules from a number of catalogs that have been marked up by hand. These rules are then applied to classify and segment the remaining catalogs.

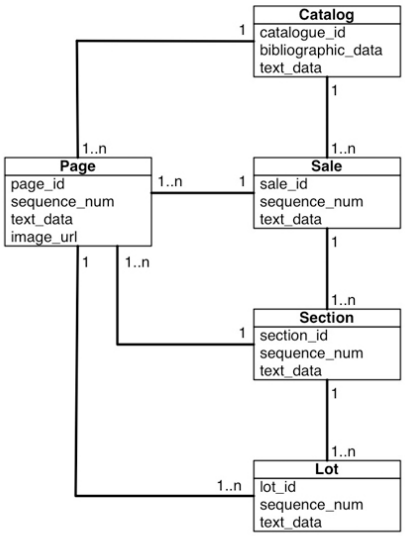
The focus of this paper is the research and creation of a system to automatically process digitised auction catalog documents, with the aim to automatically segment and label entities within and create a logical structure for each document. The catalogs processed are in an XML format produced from physical documents via an OCR process. The segmentation and assignment of entity types will facilitate, deep searching, browsing, annotation and manipulation activities over the collection. The ability to automatically label previously unseen documents will enable the production of other large scale collections where the hand labelling of the collection content is highly expensive or unfeasible.

The catalog, sale, section, lot model requires that the content of the document be distributed between these entities, which are themselves distributed over the pages of the document. Each line of text is assigned to a single entity, whole entities may be contained within other entities (a logical hierarchy), and a parent entity may generate content both before and after its child entities in the text sequence. This hierarchical organisation differentiates the problem of automatically labelling auction catalog document content from other semantic labelling tasks such as Part of Speech labelling (Lafferty et al., 2001) or Named Entity Recognition (McCallum and Li, 2003). In these tasks the classes or states can be thought of as siblings in the text sequence, rather than as having hierarchical relationships. Hence, the digitisation of auction catalog documents may require a different set of procedures to that

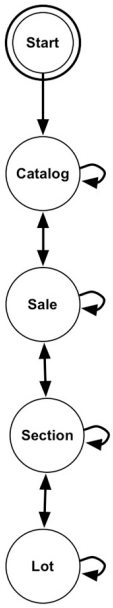
applied to, for example, the digitisation of Magazine collections (Yacoub et al., 2005) or scholarly articles (Lawrence et al., 1999).

Although a particular document model is assumed throughout this work, the theory and tools detailed can be applied to arbitrary document models that incorporate hierarchical organisation of entities.

Techniques that are successfully applied to other Natural Language Processing and document digitisation tasks may be applied to this problem. Specifically, we have developed task appropriate feature extraction and normalisation procedures to produce parameterisations of catalog document content suitable for use with statistical modelling techniques. The statistical modelling technique applied to these features, Conditional Random Fields (CRFs) (Sutton and McCallum, 2007), models the dependence structure between the different states (which relate to the logical entities in the document) graphically. A diagrammatic representation of the auction catalogue document model is given in figure 1a and an example of a model topology that might be derived from it is given in figure 1b. It should be noted that CRFs are discriminative models, rather than generative models like HMMs, a property that may be advantageous when such models are applied to NLP tasks (Lafferty et al., 2001).



(a) Document data



(b) FST transition grammar

Figure 1: Document model for an auction catalog and a Simple Finite State Transducer topology derived from it

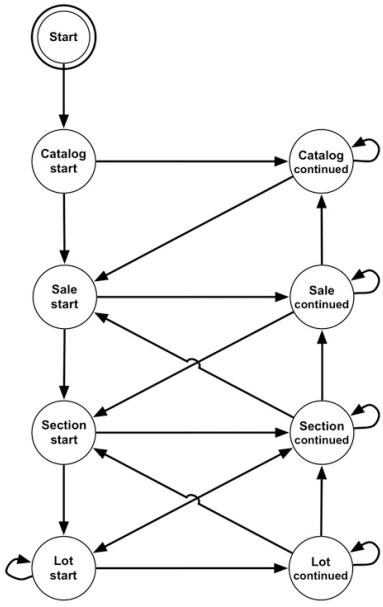


Figure 2: FST transition grammar extended to incorporate start and continue states for each entity type

The application of such techniques to hierarchically structured documents requires the logical structure of a document to be recoverable from a simple sequence of state labels. The basic transition grammar shown in figure 1b is not appropriate for this task as it is impossible to differentiate concurrent lines of a single entity from concurrent lines from two entities of the same type and relationships

between a single 'parent' entity and multiple 'child' entities. These issues may be addressed by using two states, a start and a continuation state, in the model, to represent content relating to each entity, as shown in figure 2. This modification allows the full logical structure to be recovered by post-processing the sequence of state labels.

In order to train any automated statistical learning procedure, a suitable sample of the data set must be labeled, usually by hand, to provide a target for learning or a ground-truth for evaluation. Hand labelling an XML-based representation of a document, produced via an OCR process, can be a difficult and frustrating task. To facilitate the fast labelling of documents, and thereby maximise the quantity of ground-truth data that could be produced, a cross-platform document segmentation user application was developed, shown in figure 3, that is able to format the OCR data for display and allow the user to simply and rapidly mark-up each document. This application could easily be adapted for use with other data sets and is a useful output of this work that can be used independently.

Initial experiments with the segmentation and labelling system were conducted on a hand labeled collection of 266 auction catalogs, comprising just less than 400,000 lines of text content. A detailed breakdown of the content of these documents

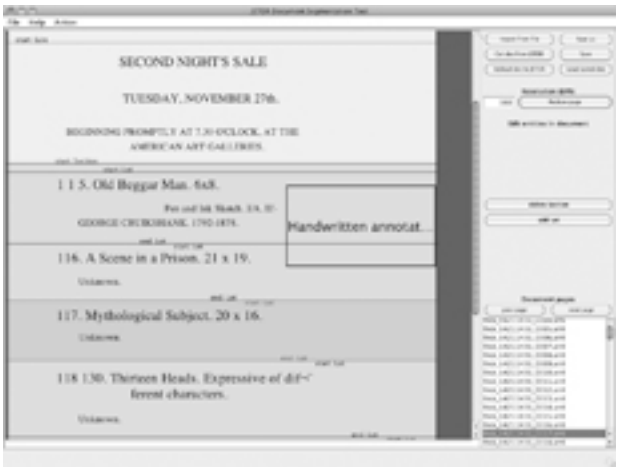


Figure 3: Document segmentation user application

is provided in table 1. These experiments compare a number of variants of the system and analyses are provided to aid in the selection of parameters for the system. The experiments are evaluated using both the retrieval statistic F-measure (Baeza-Yates and Ribeiro-Neto, 1999) and an adaption of the WindowDiff metric (Pevzner and Hearst, 2002). The system has been found to produce a satisfactory level of performance to facilitate the automated processing of auction catalog content. A breakdown of the results achieved by the system is given in table 2.

Entity	**Number of lines**
start-Catalog	266
cont-Catalog	41,275
Catalog	***41,541***
start-Sale	556
cont-Sale	3,469
Sale	***4,025***
start-Section	4,109
cont-Section	19,415
Section	***23,524***
start-Lot	91,240
cont-Lot	229,051
Lot	***320,291***
Total	**389,381**

Table 1: Number of lines corresponding to each entity type

Analysis of the errors produced by the system show that, owing to the hierarchical nature of the document model, a small number of errors at critical points in the content can lead to a large number of subsequent, concurrent errors. Unfortunately, these critical points in the ground-truth sequences are represented by the smallest quantities of content in the collection and therefore may form the weakest parts of the model. However, this also means that a small number of corrections at these key points could enable the correction of a much larger number of errors in the output. This is a useful property, as one of the design goals of this system is to facilitate the integration of feedback from users of the digitised documents into the statistical model and thereby allow the segmentations produced to improve over time.

The feedback is used to improve the model by preserving any confirmations or corrections, made by users, of the segmentation output from CRF model. This preservation is achieved by reapplying the model to the document, which has been partially labeled by users, and forcing it to pass through the user indicated states as it determines a new sequence of state labels. This allows a user to supply corrections of major segmentation errors, such as a missing high-level entity (e.g. a sale), or minor errors,

such as a single mislabeled line belonging to a lot. A user supplied correction to a major segmentation error could correct the labelling of many lines, for example by opening the Sale at the correct point and allowing the model to estimate weights for Sections and Lots thereafter, whereas minor corrections may be simply preserved so that they don't need to be reapplied to the re-segmented document.

Metric	Score
WindowDiff	0.103
F1 cont-Catalog	0.709
F1 start-Lot	0.870
F1 cont-Lot	0.934
F1 start-Sale	0.444
F1 cont-Sale	0.347
F1 start-Section	0.483
F1 cont-Section	0.548

Table 2: Results achieved by the CRF-based system calculated over 5-fold cross-validation of the hand-labelled dataset

References

Baeza-Yates, R., Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison-Wesley Publishing Company.

Lafferty, J., McCallum, A., Pereira, F. (2001). 'Conditional random fields: Probabilistic models for segmenting and labeling sequence data'. *Machine Learning-International Workshop then Conference*. Citeseer, pp. 282–289.

Lawrence, S., Giles, C., Bollacker, K. (1999). 'Digital libraries and autonomous citation indexing'. *IEEE computer*. **32(6)**: 67–71.

McCallum, A., Li, W. (2003). 'Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons'. *Seventh Conference on Natural Language Learning (CoNLL)*.

Pevzner, L., Hearst, M. (2002). 'A critique and improvement of an evaluation metric for text segmentation'. *Computational Linguistics*. **28(1)**: 19–36.

Sutton, C., McCallum, A. (2007). *An Introduction to Conditional Random Fields for Relational Learning*. *Introduction to statistical relational learning*. Pp. 93.

Yacoub, S., Burns, J., Faraboschi, P., Ortega, D., Peiro, J., Saxena, V. (2005). 'Document digitization lifecycle for complex magazine collection'. *Proceedings of the 2005 ACM symposium on Document engineering*. ACM New York, NY, USA, pp. 197–206.