Table Extraction Using Conditional Random Fields

David Pinto, Andrew McCallum, Xing Wei, W. Bruce Croft
Department of Computer Science
University of Massachusetts
Amherst, MA 01002 USA
{*pinto, mccallum, xwei, croft*}*@cs.umass.edu*

Abstract

The ability to find tables and extract information from them is a necessary component of data mining, question answering and other information retrieval tasks. Documents often contain tables in order to communicate densely packed, multi-dimensional information. Tables do this by employing layout patterns to efficiently indicate fields and records in two-dimensional form. Their rich combination of formatting and content present difficulties for traditional language modeling techniques, however. The poster presents the use of conditional random fields (CRFs) for table extraction, and compares them with hidden Markov models (HMMs). Unlike HMMs, CRFs support the use of many rich and overlapping layout and language features, and as a result, they perform significantly better.

1 Introduction

Information in many documents is carried not only in their stream of words, but also by the layout of those words [1]. Tables are often used to compactly communicate information in fields and records. They have been described as "databases designed for human eyes." They are used in everything from government reports, to magazine articles, to academic publications.

In previous work building the QuASM system (Question Answering Using Semi-structured Metadata)[6], extracting table data proved to be the most challenging task faced. Question answering (QA) systems typically perform a two-step retrieval in finding the information relevant to a query. Documents are retrieved that match the question, then those documents are searched for possible answers. QuASM used a heuristic program to turn each data cell and its meatadata into its own small document. It was found that the amount and quality of the metadata was extremely important to the success of the system, especially in the document retrieval step. Unfortunately, QuASM tended to extract too much information from the tables, which hurt performance.

To improve the extraction a move toward a more probabilistic model was proposed. The poster presents a model of table extraction that integrates evidence from both content and layout by using *conditional random fields* (CRFs) [2]. CRFs are discriminatively-trained undirected graphical models that have great freedom to use complex, overlapping and non-independent feature sets that operate at multiple levels of granularity and multiple modalities. We describe a method that simultaneously locates tables in plain-text government statisical reports, and labels each of their constituent lines with tags such as header, data, etc. The features measure aspects of the input stream such as the percentage of alphabetic characters or the presence of regular expression matching months or years.

2 Table Extraction

Table extraction may be broken down into six overlapping sub-problems:

1. Locate the table.
2. Identify the row positions and types.
3. Identify the column positions and types.
4. Segment the table into cells.
5. Tag the cells as data or headers.
6. Associate data cells with their corresponding headers.

This research concentrates on items one and two by employing a probabilistic approach to labeling the lines of a document. The labeling task can be defined as a Markov process. For example, titles are expected to be followed by table headers, followed by data rows. The clues for the probability of a label naming a line need to be derived from features of the text. This is a classic hidden Markov process where states are inferred from data observations.

3 Conditional Random Fields

Conditional Random Fields [2] are undirected graphical models used to calculate the conditional probability of values on designated output nodes given values assigned to designated input nodes. In the special case in which the designated output nodes of the graphical model are linked by edges in a *linear chain*, CRFs make a first-order Markov independence assumption among output nodes, and thus correspond to finite state machines (FSMs). In this case CRFs can be roughly understood as conditionally-trained hidden Markov models. CRFs of this type are a globally-normalized extension to *Maximum Entropy Markov Models* (MEMMs) [4] that avoid the *label-bias problem* [2].

4 Table Extraction with CRFs

4.1 Line Labels

Our approach to table extraction starts by attempting to label each line of a document with a tag that describes that line's function relative to tables. A good labeling accomplishes two goals; it marks the boundaries of the tables (table location) and identifies the row types.

Non-extraction labels represent lines that do not contribute information about table cells. The three labels are NONTABLE, BLANKLINE and SEPARATOR. Header labels mark lines that contain metadata for data cells. The header labels are TITLE, SUPERHEADER, TABLEHEADER, SUBHEADER and SECTIONHEADER. Data row labels mark rows that contain data cells, the atomic information to extract. The data row labels are DATAROW and SECTIONDATAROW. Caption labels mark rows that appear after a table but still apply to the table. The caption labels are TABLEFOOTNOTE and TABLECAPTION.

4.2 Feature Set

Pinto, et al. [6] describes features used by a heuristic table extractor to identify a narrower range of line types. This work serves as a starting point for the features used to train the Markov parameters used for labeling. One difference, however, is the treatment of separator characters, such as dashes. During the creation of a character alignment graph, these characters are treated as white space. The belief that separator characters convey a different meaning than white space prompted their inclusion in these experiments. As the system was developed other features were added to improve performance on evaluation data.

White space is employed in documents and tables to improve readability. The white space features are Four consecutive white space characters, four space indents, two gaps (where a gap is two spaces), large gaps (five spaces), single space indents, blank lines and the percentage of white space.

Text features, made up of printable characters also convey information about the type of line being observed. The text features are, three cells, header features (strings commonly found in header lines), percent of alphabet characters (A-Za-z) and the percent of digit characters.

Punctuation marks are often used for formatting tables. A line of dashes may delineate the header section from the data section of a table. Two features look directly at punctuation. They are the percentage of separator characters (-,+,—,!,=,:,*) and four consecutive periods.

4.2.1 Feature Representation

Each feature can be represented as a binary value. A 1 indicates the presence of the feature and a 0 indicates the lack of a feature. For percentage features, a threshold is set, above which the feature will score a one. With CRFs, it is as easy to use continuous-valued features as it is to use discreet-valued ones. This enabled experiments in which the percentage features were not treated as binary features, but in which the actual percentage was used as the input.

4.3 Data Set

Documents were randomly chosen from a crawl of www.fedstats.gov performed in June 2001. This random set contained documents both with and without tables. Each line of the documents was labeled with one of the twelve labels described in section 4.1.

The first fifty-two documents labeled became the training set. These documents contain 31,915 lines of text and 5,764 table lines. The next six documents labeled were set aside as an evaluation set. These documents contain 5,817 lines of text and 3,607 table lines. Finally, 62 more documents were labeled, containing 26,947 lines of text and 5,916 table lines, and were held out as final test data.

5 Experimental Results

5.1 Hidden Markov Model Training

Hidden Markov Models (HMM) are probabilistic functions of a Markov process. An HMM is specified by five parameters (S, K, Π, A, B). Since HMMs have been well discussed in the literature, [7] this section will explain how the three parameters (π, A and B) of the HMM are calculated.

The parameters π and A are calculated using a smoothed maximum likelihood estimate. The emission transition distribution is calculated using the Multi-variate Bernoulli Model [5]. In the multi-variate Bernoulli event model, a line is a binary vector over the space of features. Dimension t of the vector of line is, which is either 0 or 1, indicating whether feature occurs. We assume the probability of each feature occurring in a line is independent of the occurrence of others features in a line. We can calculate Bayes-optimal estimates for these probabilities by counting of events, supplemented by the Laplacean prior to avoid probabilities of zero or one.

5.2 Training CRFs

Using our Java implementation [3], we trained two versions of the CRF by L-BFGS. One version used only binary features (CRF Binary), and converged in 100 iterations. Another version used the actual value for the percentage features (CRF Continuous), but all other features were treated as binary, converging in 76 iterations. Both versions used a hyperbolic prior and the same conjunction of features.

5.3 Evaluation

During development of the system, a small evaluation set was used to see if performance was improving. A larger set of data was held out for final testing, and no evaluations were done on that

Data Set	HMM	CRF Binary	CRF Continuous
Training	89.2	99.7	99.7
Evaluation	85.9	95.6	96.1
Held Out	65.4	93.6	94.8

Table 1: Percent of Lines Labeled Correctly

set until the best trained CRF was decided upon. The results for the evaluation and held out test data are presented in Table 1. All three Markov processes do well on the evaluation set, but the superiority of the CRF shines through on the held out data.

6 Conclusions

Tables are formatted in many shapes and sizes. To successfully extract cell data in context requires a system that recognizes many different forms of tables, and the way the different parts the table fit into that form. This work shows that conditional random fields are especially well-suited for this task, as they can simultaneously model content and layout. As our experiments show, in all cases CRFs outperform HMMs at labeling the lines of a table.

7 Acknowledgments

This work was supported in part by the Center for Intelligent Information Retrieval, the Central Intelligence Agency and the National Science Foundation under NSF grant #EIA-9983215, the Advanced Research and Development Activity under contract number MDA904-01-C-0984 and in part by SPAWARSYSCEN-SD grant numbers N66001-99-1-8912 and N66001-02-1-8903. Any opinions, findings and conclusions or recommendations expressed in this material are the author(s) and do not necessarily reflect those of the sponsor.

References

[1] M. Hurst. *The Interpretation of Tables in Texts.* PhD thesis, University of Edinburgh, School of Cognitive Science, Informatics, 2000.

[2] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. ICML*, pages 282–289, 2001.

[3] A. McCallum. Mallet: A machine learning for language toolkit. http://www.cs.umass.edu/~mccallum/mallet, 2002.

[4] A. McCallum, D. Freitag, and F. Pereira. Maximum entropy Markov models for information extraction and segmentation. In *Proc. ICML 2000*, pages 591–598, 2000.

[5] A. McCallum and K. Nigam. A comparison of event models for naive bayes text classification. In *AAAI-98 Workshop on "Learning for Text Categorization"*, 1998.

[6] D. Pinto, W. Croft, M. Branstein, R. Coleman, M. King, W. Li, and X. Wei. Quasm: A system for question answering using semi-structured data. In *Proceedings of the JCDL 2002 Joint Conference on Digital Libraries*, pages 46–55, 2002.

[7] L. R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In A. Weibel and K.-F. Lee, editors, *Readings in Speech Recognition*, pages 267–296. Morgan Kaufmann, 1990.