Why Table Ground-Truthing is Hard*

Jianying Hu† Ramanujan Kashi† Daniel Lopresti‡
George Nagy§ Gordon Wilfong‡

† Avaya Labs ‡ Bell Labs
Avaya Inc. Lucent Technologies, Inc.
600 Mountain Avenue
Murray Hill, NJ 07974 USA

§ Department of Electrical, Computer, and Systems Engineering
Rensselaer Polytechnic Institute
Troy, NY 12180 USA

Abstract

The principle that for every document analysis task there exists a mechanism for creating well-defined ground-truth is a widely held tenet. Past experience with standard datasets providing ground-truth for character recognition and page segmentation tasks supports this belief.

In the process of attempting to evaluate several table recognition algorithms we have been developing, however, we have uncovered a number of serious hurdles connected with the ground-truthing of tables. This problem may, in fact, be much more difficult than it appears. We present a detailed analysis of why table ground-truthing is so hard, including the notions that there may exist more than one acceptable "truth" and/or incomplete or partial "truths."

1. Introduction

In past papers, we have examined the problem of table recognition and proposed algorithms we believed would be effective, at least for certain kinds of input [2]. Our goals were to address two related issues: table detection (finding the tables on a page) and table interpretation (understanding the logical structure of a table). As is common practice in the field, we began by testing our methods on small datasets we had created specifically for this purpose. Since our techniques are designed to work in both the ASCII and image domains, we built two such datasets, each consisting of a few dozen documents.

Despite the fact that we had chosen relatively "easy" tables to start off with, constructing the ground-truths proved to be time-consuming. We also discovered that occasionally the human truthers would differ in their opinions about a given table. In some cases, these alternative interpretations appeared to be justifiable – no one viewpoint was obviously better than another. In other cases, closer analysis revealed that a mistake had been made. Nevertheless, we saw no reason to suspect there might be fundamental barriers to constructing ground-truths for a much larger, more realistic dataset of tables.

After our initial success, we wanted to try a bigger experiment involving a standard document collection. We chose the University of Washington I CD-ROM (UW1), which contains 979 pages scanned mostly from journal articles [5]. From the ground-truth files provided with the images, we were quickly able to determine that there are a total of 135 marked table zones in UW1, distributed over 110 pages. We were only mildly disappointed to find that the ground-truth did not supply an interpretation for the tables. We still believed it would be easy, albeit tedious, to truth the tables ourselves, using tools we had developed for that purpose.

What we discovered in taking a deeper look at this has implications for ground-truthing not only tables, but other, higher-level document analysis tasks as well. Our main conclusion is that while it may not be surprising that table understanding is difficult for machines, it poses an extremely challenging problem for humans as well. Since possessing ground-truth is regarded as a fundamental first step towards a disciplined approach to building robust systems, this issue is a critical one.

2. Goals and Issues in Ground-Truthing

An overview of the subject of automated performance evaluation can be found in [4], which describes five fundamentally different approaches to ground-truthing (or to eliminating the need for explicit ground-truth). A survey specific to table understanding appears in [3].

Some fundamental tenets in ground-truthing include:

*To be presented at the *International Conference on Document Analysis and Recognition*, Seattle, WA, September 2001.

- The essence of ground-truth is that it must be a formal specification, couched in the same formalism as the output of the recognition system. Examples of the latter include bitmaps, RTF, spreadsheets, relational algebra, entity-graphs, etc. In model-driven systems, the specification is further restricted to those expressible within the chosen system.

- For validation, there must be a metric between different specifications of the same object (this can be binary, continuous, or vectorial). Since the models are generally data structures with operations, the metrics can be the cost of the operations required to transform one model instance into another.

- Both the automated system and the ground-truther can draw on information beyond the object itself, drawn from either the environment of the object or elsewhere. Ideally, such information can also be formally codified. It represents both the skill-level of the truther and the knowledge-base of the recognizer.

In anticipating the sorts of problems that might arise in attempting to ground-truth tables or other document understanding tasks, we have identified five broad categories:

Insufficient skill level of the ground-truther. In some instances, the ground-truth is generated by a "naive" observer: someone who is neither connected with the development of the recognition system under study, nor expert in the technical field the table is drawn from. These problems can by definition be resolved by a subject-matter specialist. This issue plays a large role in other document interpretation tasks, including automated summarization and information retrieval.

Restrictions in the model. Some tables may not fit the underlying logical model, or even a simple array model. In these instances it is tempting to define "table" to include only those objects that fit the model.

Inadequate interactive tools. Any interactive system for creating ground-truth introduces some obstacles, even when the interpretation of the table is clear.

Shortcomings in the automated table analysis system. Instances of tables that may break the analysis system fall in this category. There is often a tendency to eliminate such tables from consideration *a priori*.

Difficulties intrinsic to the table. These are caused by poor syntax or semantics, or mutilation in the composition, printing, or scanning process.

3. The Table Recognition Problem

Figure 1 illustrates the terminology we use in this paper, derived from the formalism put forth by Wang [6]. At the lowest level, a table is composed of two types of cell: the *Dcell*, or data cell, and the *Acell*, or access cell. The former comprise the core of the table, while the latter occur within headers. In the example, the data cells are all numeric (stock prices), while the access cells are all alphabetic. Working upwards in the hierarchy, these basic cells are organized into columns and rows. The column headers are grouped into a region named the *Box*, and the row headers into a region called the *Stub*. The header for the box/stub (if there is one) is known as the *Stub Head* or *Box Head*. The collection of all the data cells makes up the *Body* of the table.

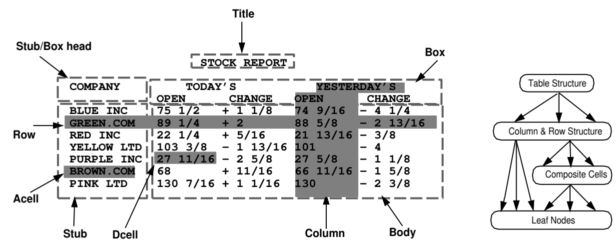

Figure 1. Table terminology (adapted from Wang's Ph.D. thesis).

While it is traditional to regard document analysis results as tree-structured, we have adopted a slightly more general representation, a directed acyclic graph (DAG). There are two basic classes of nodes in our table DAG: *leaf nodes* which have no children and contain content corresponding to a specific region on the page (*i.e.*, one or more text strings), and *composite nodes* which are simply unordered collections (sets) of previously-defined leaf and composite nodes. Edges encode the *contains* relationship. While every node has an optional label, there is no rigid policy enforcing how nodes must be labeled relative to one another.

To enable the viewing of document analysis results and to support the ground-truthing process, we have developed an interactive tool we call **daffy** for browsing and editing table DAG's. The user interface portions of the system are written in Tcl/Tk. Input is accepted in both image (TIF) and text (ASCII) formats. The full generality of the graph model described in the preceding section is supported. Consistency of the graph is maintained automatically without placing burdensome constraints on the user.

4. Empirical Evaluation of the Table Ground-Truthing Task

As with other higher-level document understanding tasks, the meaning of any table is open to interpretation. There is, of course, an authoritative reading for every one of the tables in the UW1 dataset: that which the author intended. In practice, however, the real question is whether the ground-truthing can be performed successfully under a reasonable set of assumptions, or whether there are barriers that impede or even prevent us from achieving this goal.

One perspective on this issue can be found by studying users' actual attempts to ground-truth table data. A sensible question to ask is, can technically-adept users, sharing a common understanding of the problem at hand, arrive independently at ground-truths that are largely the same, or will there be significant differences? In [4], Nagy notes "It is clear that the reference data must be labeled at least an order of magnitude more accurately than the expected accuracy of the system to be tested." To examine the inherent difficulty in ground-truthing tables from UW1, we designed a simple experiment involving five representative examples:

a049 *TABLE 3: Concentration of ^{210}Po in Vegetables and their Associated Soils at the Control Site.* This table has three logical dimensions (*Plant*, *Organ*, and *Radionucleotide Concentration*), as well as multi-level row headers. Either the whole plant (*e.g.*, *Grass*) or parts of it are measured. Its interpretation is relatively straightforward, although one might require domain knowledge in botany to understand it.

c003 *Table 1: The domain connectivity matrix IX in the example.* This table is actually a matrix of graph connectivity information for a finite element analysis. Matrices fit the Wang model only if column and row indices are added as headers, so that they can be permuted without loss of information. Hence, the physical structure is clear, but the logical structure is implicit in the layout.

h00u *TABLE 4: Results of U.S. Pharmacopoeia (1990) and B. Pharmacopoeia (1988) trials on two Spanish bentonites.* This table is folded to conserve space, so that column headers are interspersed with data cells. While the logical structure is simple, the folding complicates the physical layout.

h01c *Table 1: Example of traditional station adjustment procedure and error analysis for horizontal direction measurements.* This rather dense table shows four pairs of relative angular observations (deg, min, sec), in each pair one is for the telescope reversed. Various statistics are embedded in (and interrupt) the column and row structure. The headers do not completely span their subordinate columns, making it difficult for a non-specialist to know how to associate some columns of numbers.

i04l *Table 1: Salient features of nuclear refinery plant concept.* This table appears to have been designed more with space-efficiency than readability in mind. It contains overlapping cells with varying amounts of text; segmenting the cells requires some effort as well as domain knowledge.

All of the page images were pre-processed using Baird's **pagereader** system to identify word bounding boxes [1]. Four of the authors of this paper were then each instructed to ground-truth the tables following a predetermined, bottom-up procedure. When in doubt, the ground-truthers were told to use their discretion.

The truthers were also asked to keep track of how long it took, and to record responses to the following two questions: (1) Did you complete the ground-truthing for the table? and (2) How confident are you that you ground-truthed the table correctly? The task took from six minutes to half an hour per table. The ground-truthers all had some tables they were uncertain about, the consensus seeming to reflect that a049 was the easiest and h01c and i04l the hardest.

Once the ground-truthing was complete, the resulting table graphs were compared using *graph probing*, a general paradigm for quantifying the similarity between two arbitrary directed acyclic graphs. Conceptually, the idea is to place each graph inside a "black box" capable of evaluating a set of graph-oriented operations (*e.g.*, returning a list of all the leaf nodes, or all nodes labeled in a certain way). We then pose a series of probes and correlate the responses of the two systems. A measure of their similarity is the number of times their outputs agree (see [2] for further details). Note that since the series of probes are generated using one of the graphs (called the *probe source*), this similarity measure is not symmetric. Our goal is to measure the consistency across different ground-truthers. The results for this experiment comparing the graph probing agreement for the graphs produced by the four truthers are given in Table 1.

	Ground-	*Probe Source*			
Table	*Truther*	*A*	*B*	*C*	*D*
a049	*A*	–	0.87	0.88	0.84
	B	0.81	–	0.93	0.90
	C	0.82	0.94	–	0.85
	D	0.78	0.90	0.85	–
c003	*A*	–	0.96	0.99	0.98
	B	0.96	–	0.97	0.96
	C	0.98	0.95	–	0.98
	D	0.96	0.93	0.96	–
h00u	*A*	–	0.49	0.67	0.64
	B	0.51	–	0.56	0.66
	C	0.53	0.43	–	0.75
	D	0.49	0.45	0.66	–
h01c	*A*	–	0.77	0.44	0.74
	B	0.70	–	0.61	0.89
	C	0.41	0.62	–	0.63
	D	0.66	0.90	0.61	–
i04l	*A*	–	0.82	0.78	0.86
	B	0.67	–	0.79	0.85
	C	0.66	0.84	–	0.78
	D	0.71	0.86	0.80	–

Table 1. Results for the table ground-truthing experiment (graph probing agreement).

In examining the results, we see that for table a049 there is good agreement among the four ground-truthers. Hence, this table is relatively "easy." While table c003 was anomalous in that it did not possess the logical structure required of the Wang model, it generated the highest agreement in our study, a nearly perfect consensus. (Recall, though, that **daffy** implements a more flexible, lower-level model than the full Wang model.)

Test table h00u, on the other hand, presents an enormous contrast. While this may be a "simple" table logically once it has been unfolded, there was substantial disagreement among the ground-truthers. Note that this cannot simply be a function of some truthers regarding the table as folded, and others regarding it as not. If that were the case, there would be agreement between at least some subset of the ground-truthers.

There was also a fair amount of disagreement for table h01c, but less so than in the previous case. Examining the graph probing scores, it appears as though Truther C arrived at a relatively "unique" interpretation, quite distinct from the other three. Truthers B and D were of like minds for the most part. Finally, table i04l demonstrated moderate agreement. Truther A seems to have produced a more detailed truth than the others (this would explain why the column of scores generated using Truther A's mark-up is noticeably lower than the rest).

5. Further Observations on the UW1 Tables

We analyzed all of the 135 tables in the UW1 dataset and found that 83 of them were anomalous for one reason or another. Not all of these cases contained "hard" logical structure. Still, all of these cases are likely to defeat standard document analysis systems not prepared to cope with the myriad special cases that arise in even a modest-sized dataset like UW1. We characterized the anomalies into 16 categories and present our analysis of the "hard" UW1 tables in Table 2.

Category	*Page Image Containing Table Instances*
Folded Tables	d06f, h00t, h00u, h046, s00d
Implicit Headers	c005, h01c, j04p, n04e, v00k, v00l, v00n
Multi-Level Headers	a046, a048, a049, a04h, a05f, d039, d06f, d06m, e00d, e011, e012, e02e, e02j, h00t, h015, h031, h035, h03e, h044, h049, j03e, j04p, n038, n039, s011, s012, s02e, s02j, v00l
Overly Complex Structure	a04g
Graphics Symbols	c008, c03i, v002
Mathematics	a002, a036, d04b, e04g, h015, h03e, h049, s04g
Matrices	c003, c007, d067
Header References	j03l
Variable Column Structure	d04b, n02d, v00d
Folded Lists	n05c
Ambiguous Structure	a036, c007, d040, d05d, e03l, h01c, i005, i04d, i04l, ig05, n04b, s03l, v00c, v00k
Specialized Domains	a05f, a05h, d05d, d067, n02d, n037
Vertical Row Header	h044
Sideways	a002, d067, e03f, h008, h00t, i032, s03f, v009
Implicit Table	j00l
Multi-Line Cells	a061, c036, c039, d040, d04b, e011, e012, e03f, e04g, h035, h03c, h03f, i049, i04b, i04l, k007, k008, k00j, n01n, s011, s012, s03f, s04g, v002, v009, v00c, v00d

Table 2. Categories of anomalous tables and their instances in the UW1 dataset.

Folded tables. These are tables that are folded vertically or horizontally (with duplicated column/row headers) because the original tables are too tall or too wide to fit in a given space. In order to ground-truth such tables, the system needs to provide mechanisms to merge rows or columns (that have been broken in the folding process) as well as to handle the duplicated headers.

Implicit headers. Sometimes the headers of certain rows or columns are omitted, presumably because they are implied by the topic of the document or the surrounding text. Such headers are nonetheless important to extract in order to summarize or query the table content. Should they be inferred and recorded as part of the ground-truth? If so, how to handle ground-truth that may not be unique?

Multi-level headers. Multi-level headers are structured in a hierarchical manner, with the domain defined by a higher-level header applying to all its children. This is a common technique for displaying tables with more than two logical dimensions and is covered by the Wang model. However, ground-truthing such tables is non-trivial and particularly prone to error because (1) the relations among multiple levels of headers are often difficult to identify; (2) the interface for annotating many levels of hierarchy among objects laid out in two dimensions is inherently complex.

Overly complex structure. An example of this is a04g. Instances of these tables (a04g) are very difficult to understand because it involves multiple levels of indexing structured in an unconventional way that cannot be formulated using the Wang model. In some sense, it is a complex combination of multiple tables. In order to ground-truth such a table, one would need to come up with a table model that is even more sophisticated than the Wang model.

Graphics symbols. It is not clear how cells containing graphics symbols should be recorded – record the symbol directly as an attached image, or encode the semantics of the figures, or both?

Mathematics. When the content of a table cell is a math equation, the system must allow some kind of script (*e.g.*, LaTeX) to enter it. The implication is that it is not sufficient to use plain ASCII to encode cell content.

Matrices. Some zones in UW1 labeled as tables are really matrices, with no headers and no apparent logical relation between the cells. While these can be considered tables with implicit headers, inferring the meaning of the 2-D layout could require a thorough understanding of the surrounding text.

Header references. In one of the tables (j03l), the column headers are indices which point to an expanded list of headers in what would normally be considered the footnote area. It is not clear how such referencing mechanisms should be represented – is it sufficient simply to record the expanded headers as the "true" headers?

Variable column structure. In some tables certain data cells expand across multiple columns, interrupting the "normal" column structure of the table. This seems to be used mostly to accommodate certain "abnormalities" in the data and induces yet another level of structural complexity not covered by the Wang model.

Folded lists. Sometimes a region is classified as a table because of the regular, grid-like layout, but a close examination reveals that there is no logical structure in place and it is really just a folded list (n05c). Such regions perhaps should not have been labeled as a table in the first place.

Ambiguous structure. Sometimes it is simply impossible to decipher the logical structure of a table due to misalignment (h01c), missing headers (h01c), complex structure within cells (i005), or just poor layout (v00c). Should one simply give up in these cases, or try to come up with some "intermediate level" ground truth which only attempts to record the physical structure?

Specialized domains. Certain tables in highly specialized domains require domain-specific rules to interpret their content. UW1 in particular includes a number of tables containing chemistry symbols and formulae (n02d).

Vertical row headers. One table has a high-level row header printed vertically (h044).

Sideways. Tables printed sideways to fit on the page.

Implicit table. Tables that are not labeled as such in the original document (j00l).

Multi-line cells. These tables are not necessarily difficult from a logical standpoint, but understanding how to properly segment the cells may require natural language understanding or specific domain knowledge (i041).

6. Discussion

We have sought to identify some of the fundamental issues that make ground-truthing tables hard. These same questions are likely to arise in other high-level document understanding applications. When there is significant disagreement between ground-truthers, as we have demonstrated, is it the fault of the table? The ground-truthers? The table model we are trying to use? The tool used to create the mark-up? The graph comparison metric? Or all of the above? Wherever the "blame" might lie, ultimately all of the above components must function together as a complete system in order for ground-truthing to be successful.

Appealing to subject matter specialists for help in interpreting our tables and attempting to define more powerful table models could perhaps ameliorate some of the difficulties we have seen, as might developing better tools. The **daffy** interface we use for ground-truthing is very flexible. This makes it easy to use, but also gives the truthers an opportunity to mark-up the same table in a number of different ways. Could one build a **daffy**-like tool that forces ground-truthers to adopt a consistent interpretation? There is a tradeoff here, however. Making the tool more rigid would make it harder to use, and it is unclear whether such an approach could offer any guarantees.

Even the notion that there is a single, well-defined ground-truth is open to debate. If we were to allow multiple, competing ground-truths, how would that effect performance evaluation? Is it possible to know *a priori* how many truths are required for a given input?

It is certainly possible to imagine truthing applications where certain aspects of the document structure are easy to discern while others are much more vexing. Should the ground-truthers be forced to complete the mark-up in every instance, or is there a way to allow for partial ground-truthing? But, again, what is the proper way to handle evaluation under such circumstances?

References

[1] H. Baird. Anatomy of a versatile page reader. *Proceedings of the IEEE*, 80(7):1059–1065, 1992.

[2] J. Hu, R. Kashi, D. Lopresti, and G. Wilfong. Table structure recognition and its evaluation. In *Proceedings of Document Recognition and Retrieval VIII*, volume 4307, pages 44–55, San Jose, CA, January 2001.

[3] D. Lopresti and G. Nagy. A tabular survey of automated table processing. In A. K. Chhabra and D. Dori, editors, *Graphics Recognition: Recent Advances*, volume 1941 of *Lecture Notes in Computer Science*, pages 93–120. Springer-Verlag, Berlin, Germany, 2000.

[4] G. Nagy. Document image analysis: Automated performance evaluation. In A. L. Spitz and A. Dengel, editors, *Document Analysis Systems*, pages 137–156. World Scientific, Singapore, 1995.

[5] I. Phillips, S. Chen, and R. Haralick. CD-ROM document database standard. In *Proceedings of Second International Conference on Document Analysis and Recognition*, pages 478–483, Tsukuba Science City, Japan, October 1993.

[6] X. Wang. *Tabular abstraction, editing, and formatting*. PhD thesis, University of Waterloo, 1996.