ATTN: Towards Practical Automated Tabular Semantic Analysis

Kan Chen, Teck Wei Low, and Alex Q. Chen

Infocomm Technology Cluster, Singapore Institute of Technology, 10 Dover Drive, Singapore

ABSTRACT

Access to printed copies of documents is only available in many organisations due to legal restrictions. Digitalising these documents has several challenges, such as overlapping texts and cancellations due to manual editing, varying layouts, low contrast, physical damages, and high cost for cloud-based (e.g., AWS) bulk processing. This paper introduces a low-cost practical method for analysing tabular semantics in printed document digitisation. We propose to first extract the text labels followed by text values and table structure semantics, then refine the extraction. Our method leverages Fuzzy matching, and Spatial hashing to facilitate the extraction. The results showcase that our method is effective and efficient with less than 1 cent/page cost on AWS.

Keywords: Tabular, Semantic Table Interpretation, Image processing, AWS

1. INTRODUCTION

In numerous organizations, access to documents is limited to printed copies, a situation shaped by historical factors and legal constraints. These paper-based documents are often the only available records. To bring these documents into the digital age, it's necessary to digitize both printed and handwritten content, transforming them into electronic formats.

This process, however, presents significant challenges. Manual alterations and the poor quality of text in these documents frequently lead to difficulties in accurately extracting information. The result is often a digital copy that is incoherent or contains errors, complicating the digitization effort.

To provide meaningful abstractions of information such as forms and table documents, various techniques have been explored, from rule-based algorithms to Artificial Intelligence (AI)/Machine Learning (ML) based solutions to extract tabular data. However, existing rule-based methods usually require crafting many heuristics, which is often not easy. AL/ML solutions usually require cumbersome cleaning, dataset preparation, and AI/ML model training, which would require extensive expertise and effort. Moreover, bulk processing vast amounts of documents (e.g., in AWS cloud settings) they are still expensive. We aim to achieve this: the performance is less than 1 second of processing time per page, and the cost is expected to be less than 1 cent per page.

In this paper, a novel low-cost rule-based algorithm method for Automatically Tabular Semantics Analysis (namely, ATTN) is proposed. It has the following features.

- Unlike some existing methods focusing on the appearance of the data, ATTN focuses on the table semantic data. As JavaScript Object Notation (JSON) is the popular data format to represent data, we propose one JSON-like representation (label-value) to represent the tabular semantics hierarchically. Moreover, the location information is also considered.
- Different to existing methods that usually require a tedious training process or formulating complex rules, ATTN is low-cost and lightweight. We perform the in an unsupervised manner by checking the table and entity self-similarity and self-regularity.
- The simplicity of ATTN allows it to be easily integrated into a cloud platform like AWS and be a practical solution for table contents analysis.

Kan Chen, Teck Wei Low, and Alex Q. Chen are with Infocomm Technology Cluster, Singapore Institute of Technology
Send correspondence to Kan Chen and Alex Q. Chen. E-mails: {kan.chen,alex.q.chen}@singaporetech.edu.sg

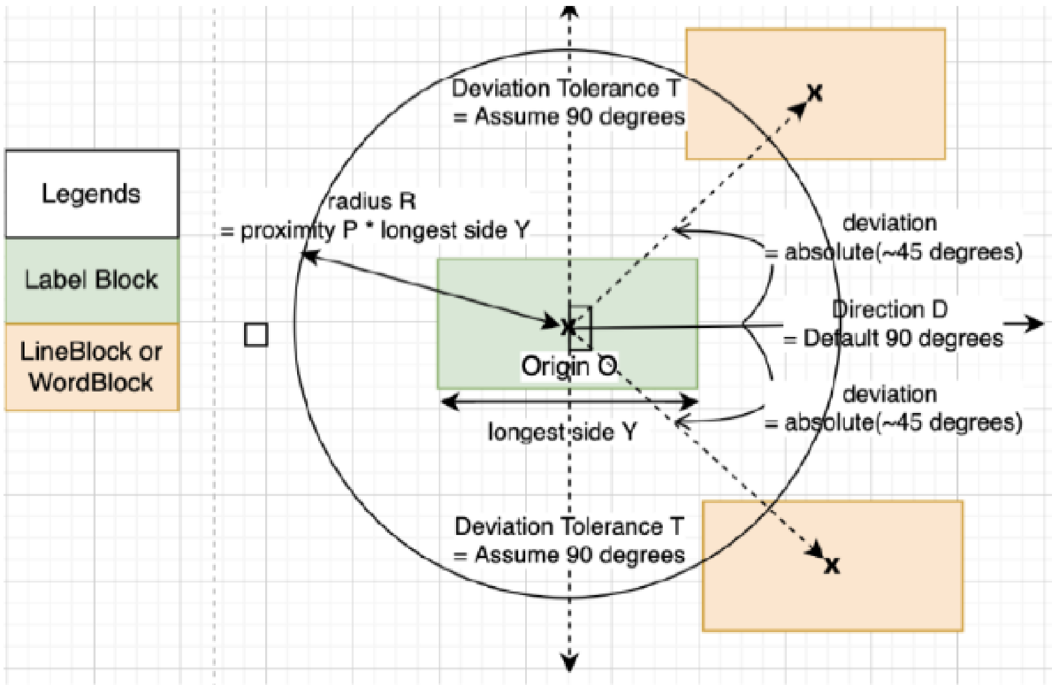


Figure 1. Extracting the values.

2. RELATED WORK

Existing market solutions (such as Rossum and Nanonets) to automate the digitalisation of data from physical mediums, like physical papers, into digital formats cost between $0.45 to $2.03 per invoice,[1] or $0.10 per page with $499 a monthly subscription fee. Amazon Web Service (AWS) provides a hosted OCR service called AWS Textract which has printed and handwritten text extraction capabilities, and it uses the pay-by-usage pricing model, which is relatively economical. The AWS Detect Document Text API service cost $0.0015/page.[2] These cost models are not financially feasible for bulk extraction for many organisations.

Techniques using deep learning models to extract table structure automatically from an input image have been shown feasible,[3],[4].[5] Deep learning methods can be implemented to understand complex table semantics.[6] Building such models often requires a lot of expertise and experimentation that can be expensive and difficult to have. Others have explored methods to analyse table structures and obtain semantic information.[7] Please refer to the survey.[8]

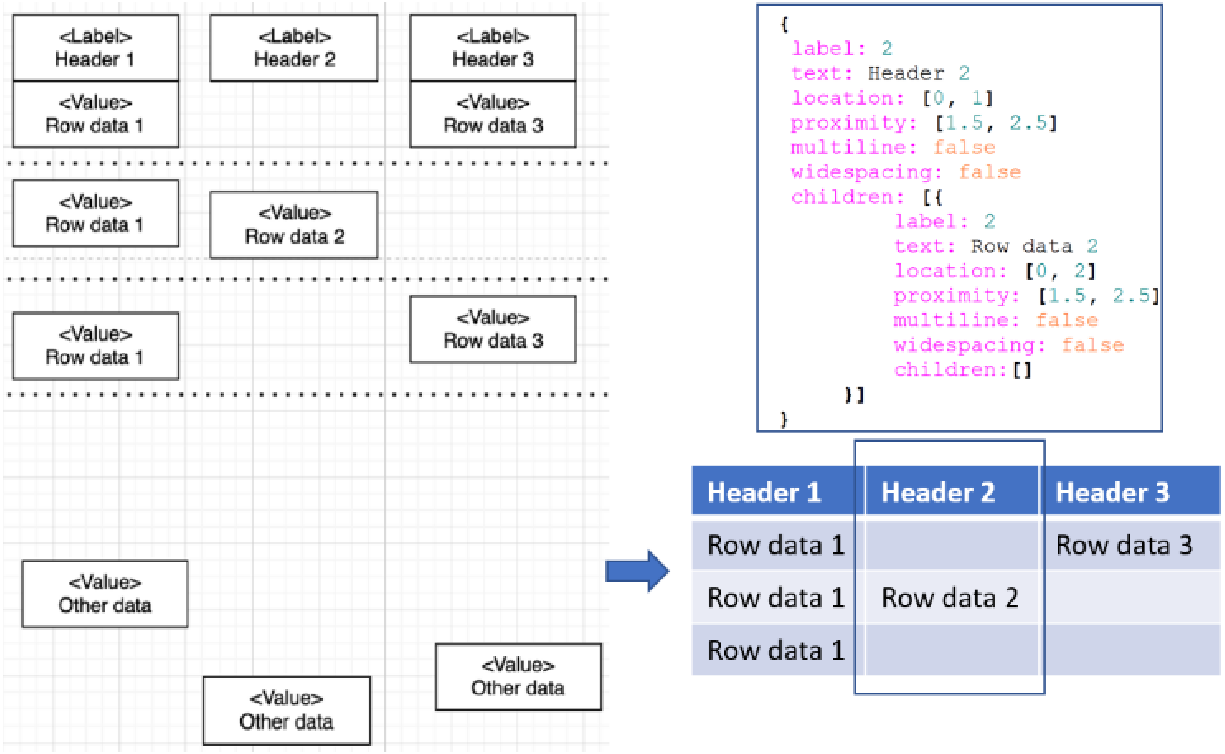
Within the range of these solutions, there's a frequent need for specific datasets, heuristics, and hardware. However, acquiring these resources is not always straightforward or feasible. The availability and accessibility of these elements can pose significant challenges, especially in resource-limited environments.

To navigate these constraints, the implementation of a lightweight and unsupervised approach holds considerable promise. Such a method would be more practical and adaptable for various applications, offering greater scalability. This approach could significantly simplify processes and expand the potential for broader application and utility.

3. METHOD

For ease of integration, AWS is the preferred ecosystem. The proposed algorithms primarily rely on generating potential template-based matches and refining.

(1) The document is firstly processed using basic OCR such as (AWS Textract) to obtain Line and Word blocks. Spatial hashing (two-dimension grids with 1s and 0s based on whether the cell is occupied) was unitised to facilitate the computation and improve the performance.



```
{
 label: 2
 text: Header 2
 location: [0, 1]
 proximity: [1.5, 2.5]
 multiline: false
 widespacing: false
 children: [{
          label: 2
          text: Row data 2
          location: [0, 2]
          proximity: [1.5, 2.5]
          multiline: false
          widespacing: false
          children:[]
      }]
}
```

Header 1	Header 2	Header 3
Row data 1		Row data 3
Row data 1	Row data 2	
Row data 1		

Figure 2. Extracting the table.

Hamming distance-based clustering is used to obtain the potential labels. If one grid cell is occupied, we assign value 1 otherwise 0 to that cell. Then the grids can be represented as a 2D bit matrix. Then if several consecutive rows are close to each other in terms of Hamming distance. Then we consider they form a table and use the first row as the header. Inside this header, based on their inter-bit block distances, we select the most common inter-bit block distance to separate the header blocks into label blocks. In addition to these labels, the user can also adjust, define, or add common labels like “Name” to create a label set for label extraction.

In addition to these labels, the user can also adjust, define, or add common labels like “Name”. This set of labels will be used for the label extraction.

(2) Labels are then refined and retrieved using Fuzzy matching based on Levenshtein Distance as the preprocessing OCR may not be exact matches due to misrecognition. It is performed recursively to extract label blocks. Both similarity scores (0.85 as the threshold in our implementation) and partial similarity scores (0.80 as the threshold in our implementation) are obtained and applied for matching.

(3) In the Line/Word blocks, the values are obtained by considering their relative position to the labels (Fig. 1). That is, the values need to be positioned near their corresponding labels on the vertical axis, maintaining a close alignment without deviating excessively (staying within a 45-degree range). This process is recursive, additional values will be determined based on the values already found.

(4) As shown in Fig. 2, the structure of the table is done by using a layered approach and leveraging on the label extraction to identify the location of the table headers. The same algorithm as step (1) for slicing labels is applied vertically to extract a two-dimensional (2D) matrix of the table. Note that several values might be combined as a single multiline value. The table will then be refined to eliminate far-away values.

Those obtained labels and values are sorted either horizontally or vertically by position to identify column or row orderings. Value extraction is then used to probe for the first column or row of the table. The rest of the columns or rows are obtained using distance-based clustering. Any rows/columns that are too far away from the previous rows/columns will signify the end of the table.

4. RESULTS

The accuracy evaluation of our methodology is executed by comparing two types of outputs. The first collection, designated as the Controlled outputs, involves the manual extraction of information from the client's documents and its subsequent organization into JSON formats, achieved through direct human intervention. Conversely, the Experimental outputs are processed using our proposed automated approach. Subsequent to the extraction process, the results are then compared in a diff-like manner.

The results on our datasets with nine documents that cannot be correctly handled by AWS Table Extraction ($0.015 per page) are as follows: An average accuracy of 93.38%, time of 0.15s, and cost of $0.0015 per page.

5. CONCLUSION

In this paper, we introduce ATTN, a novel method developed to digitize scanned printed documents. ATTN leverages Fuzzy matching combined with Spatial hashing, a synergy that enables it to reach an ideal accuracy rate of 93.38%. This method offers a cost-effective and efficient solution for converting printed material into digital format in a JSON-like structure maintaining the inherent structure and information.

Furthermore, ATTN is characterized by its low-cost and lightweight nature, making it an ideal candidate for integration into popular cloud services like AWS. This compatibility with widely used cloud platforms enhances ATTN's practicality and accessibility, allowing it to be easily adopted and utilized in various settings.

Looking ahead, our future work will focus on refining ATTN's capabilities in handling more complex table extractions. While ATTN currently is able to handle the task of digitizing standard tables, the intricate and varied layouts found in some documents pose additional challenges. Our plan is to enhance ATTN's algorithm to more effectively identify and accurately digitize these complex tables and forms. Such improvements will further extend ATTN's applicability and efficiency, making it an even more powerful tool in the realm of document digitization. With these advancements, we aim to set a new standard in the accuracy and versatility of converting printed documents into their digital counterparts.

REFERENCES

[1] Baudis, P., "The Real Cost of Invoice Automation: The TCO of Invoice Data Capture (Part 3)." https://rossum.ai/blog/the-tco-of-invoice-data-capture-cognitive-cloud-based-solution-3/. (Accessed: 19 January 2023).

[2] Amazon, "Intelligently Extract Text Data with OCR - Amazon Textract Pricing - Amazon Web Services." https://aws.amazon.com/textract/pricing/. (Accessed: 19 January 2023).

[3] Clark, K., Luong, M.-T., Le, Q. V., and Manning, C. D., "Electra: Pre-training text encoders as discriminators rather than generators," *arXiv preprint arXiv:2003.10555* (2020).

[4] Iida, H., Thai, D., Manjunatha, V., and Iyyer, M., "Tabbie: Pretrained representations of tabular data," *arXiv preprint arXiv:2105.02584* (2021).

[5] Tensmeyer, C., Morariu, V. I., Price, B., Cohen, S., and Martinez, T., "Deep splitting and merging for table structure decomposition," in [*2019 International Conference on Document Analysis and Recognition (ICDAR)*], 114–121, IEEE (2019).

[6] Kurama, V., "Form Data Extraction. Retrieved from Form Data Extractio." https://nanonets.com/blog/form-data-extraction/. (Accessed: 19 January 2023).

[7] Zhang, Z., "Towards efficient and effective semantic table interpretation," in [*The Semantic Web–ISWC 2014: 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I 13*], 487–502, Springer (2014).

[8] Liu, J., Chabot, Y., Troncy, R., Huynh, V.-P., Labbé, T., and Monnin, P., "From tabular data to knowledge graphs: A survey of semantic table interpretation tasks and methods," *Journal of Web Semantics* , 100761 (2022).