1 Description

The goal of my project will be to construct a machine learning algorithm to extract and label tables from DOM tree representations of web pages. The machine learning algorithm will be a supervised learning algorithm; we will attempt to use sequence-to-sequence learning methods commonly used in machine translation. The algorithm will take DOM trees as input and output the path of the HTML object associated with each cell of the table. The difficulty of the project comes from the irregular format of the input and output, the relatively small amount of training data available, and the lack of a clear method to evaluate the performance of our algorithms. In particular, finding effective ways to represent the input trees as sequences and to encode the output tables as sequences is a difficult subtask; this must be done in order to apply the sequence to sequence machine learning algorithms. Training data will consist of webpages and associated tables that are labeled by hand, but its collection is not a significant part of this project. To evaluate our ML algorithms, we will compare the tables it generates to the tables constructed by hand, but we will need to determine as part of this project how to evaluate the generated table if it contains, for example, swapped rows, extra columns, partially correct data in a cell, and so on.

Currently, tabular data stored on websites is often not constructed with the `<table>` HTML tag, making it difficult to automate the extraction of this data. Screen readers are able to detect the rows and columns of tables where the table is implemented with a `<table>` tag, however, previous work on the Smart Wrap project suggests that screen readers cannot efficiently parse tabular data that is not implemented with the `<table>` tag [1, 2]. In reality, tables on websites are often implemented with a complicated mix of `<div>`, `<span>`, and other such elements. If successful, this project could allow blind web users who use screen readers to obtain tabular information from websites with tables that are not implemented with the `<table>` tag. This project also has applications in data mining and question answering, as the tables from web sites contain valuable knowledge that could be hard to extract.

My project will be done in collaboration with Professor Anthony Tomasic, Lucile Callebert (postdoc), and Zhan Dong (undergraduate).
2 Project Goals

2.1 100% Goal

If all goes as expected, by the end of the semester, we will have constructed a supervised learning method to produce tables from web pages. We will also have created performance metrics to evaluate the tables output by our ML algorithms, and our ML algorithms will have reasonable performance on actual web pages using our performance metrics. We will have explored methods for encoding trees and tables as sequences. Near the end of the semester, we will write a paper summarizing our results.

2.2 75% Goal

If the project goes slower than expected, perhaps all supervised learning methods that we attempt will not perform well. We will still gain experience with methods to encode trees and tables as sequences, and our attempts at a supervised learning algorithm itself can motivate future attempts, contributing to the literature on table extraction.

2.3 125% Goal

If the project goes faster than expected, perhaps the quality of tables produced by the supervised learning method will be comparable to or even exceed human performance. Hopefully, the paper we write will be accepted and published by a conference.

3 Milestones

End of Fall Semester: I will learn about the high-level goals of the Smart Wrap project. I will have completed a significant portion of the literature search so that I can begin attempting to construct the supervised learning methods within the next two weeks.

January 31st: I will complete the literature search and begin attempting to construct supervised learning methods. A significant subtask is to find and think about effective ways of representing the input (trees) and the output (tables) in ways that are understood by supervised learning methods (generally, real numbers and categories, but perhaps the literature review will reveal other ways), so this is where I will try to begin.

February 14th I will have made significant progress on a supervised learning method.

February 28th I will have finished writing code for a supervised learning method, and will test the performance of this method. I will start attempting other methods/improving this method based on the results.

March 21st I will continue iterating and improving the performance of multiple supervised learning methods.

April 4th I will continue improving the supervised learning methods, and test them to obtain results. I will begin to write a paper summarizing the results of the project.

April 18th I will finish writing the paper to summarize my results.

May 2nd I will make edits to the paper and the supervised learning methods as required.
4 Literature Search

One main difficulty of this project is to figure out ways to represent the input trees and output tables as sequences. I will look for papers on sequence to sequence supervised learning algorithms, and in particular those that use trees or tables as input or output. I will also read papers the previous work on the Smart Wrap project; some previous students have written machine learning methods for similar tasks. Previous work on the problem of extracting tabular data from webpages is also important; I will look for research papers on this topic as well. I believe previous work has mainly focused on unsupervised learning methods, but these could still be quite useful in constructing a supervised learning algorithm.

5 Resources Needed

I will interact the the SmartWrap tool built earlier in the project; this tool is used for labeling tables on web pages. Although unlikely, it is possible that the supervised learning methods we attempt will require expensive computation; it should not be too difficult to obtain processing power through Amazon Web Services or CMU resources.

6 Website

Project information will be posted at http://www.andrew.cmu.edu/user/vzx.

References
