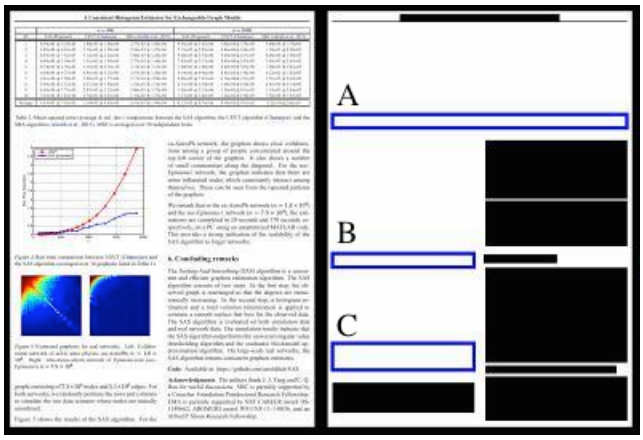

Figure 4: Disambiguating caption to empty space pairing. From the original document (left panel, page from (Aziz and others 2011)) text regions and caption regions are detected (shown as filled and empty boxes in the right panel). At this point it is ambiguous what space to assign to the middle caption, labelled as 'B', because considered in isolation this caption could plausibly refer to the region above or the region below it. However our algorithm detects that the lower caption, caption C, only has one large, empty region of space nearby that it could refer to. Once it is known that that space has to be assigned to caption C it becomes clear caption B must be referring to the region above it.

based on how likely it is to contain a figure. Our scoring function rejects regions that contain only empty pixels, or if they are too small. Otherwise regions are scored based on their size, plus a bonus if they contain a large graphical region and a smaller bonus if they contain a smaller graphical region or many different graphical regions. This heuristic helps us to be robust to errors in the previous steps. For example, if a caption has a line plot above it and a bulleted list below it (that was classified as image text), our scoring function will find it preferable to select the line plot's region over the bulleted list's region as that caption's figure region. However if no image regions are found, for example, if the caption really is about a bulleted list, the scoring function will then allow captions to be assigned to regions that contain nothing but text.

Region selection determines which of the proposed regions to select for each caption. A naive approach would be to iterate through each caption and select its highest scoring region, however this can lead to errors in the face of ambiguities. Consider Figure 4, where captions are adjacent to multiple figures making it impossible to judge which region to assign to each caption when considered in isolation. To deal with such ambiguity we iterate over all possible matchings of accepted figure regions to captions and select the highest scoring configuration. In practice the number of possible configurations is almost alway very small (less than 5) since most pages have only a few figures on them, and each figure typically only has a few proposed figure regions that do not get rejected by the scoring function. Use of this strategy for figure-caption assignment rather than iteratively assigning the highest scoring proposal to its corresponding caption increases our precision by about 2.5% and recall by about 1.5% in our analysis.

	Precision	Recall	F1
Ours	0.957	0.915	0.936
Praczyk and Nogueras-Iso	0.624	0.500	0.555
pdfimages	0.198	0.116	0.146

Table 1: Precision and recall on figure extraction.

	Precision	Recall	F1
Ours	0.952	0.927	0.939
Praczyk and Nogueras-Iso	0.429	0.363	0.393

Table 2: Precision and recall on table extraction.

A troublesome case for our region proposal method is when figures are directly adjacent to each other, in which case it is difficult to tell where one figure ends and where the other one begins. This is shown in Figure 3. In these cases proposed regions might get expanded too much and include multiple figures. To handle this problem, during the region selection stage, if we detect proposed regions would overlap, an attempt is made to split the conflicting region based on areas of whitespace inside the overlap. We found this increases our recall by about 2%.

4 Dataset

We have assembled a new dataset of 150 documents from three popular computer science conferences. We gathered 50 papers from NIPS 2008-2013, 50 from ICML 2009-2014, and 50 from AAAI 2009-2014 by selecting 10 published papers at random from each conference and year. Annotators were asked to mark bounding regions for each figure and caption using LabelMe[7]. For each region marked by annotators, we found the bounding box that contained all foreground pixels within that region. These bounding boxes were then used as ground truth labels. In total we acquired bounding boxes for 458 figures and 190 tables. Our dataset along with the annotations can be downloaded at pdffigures.allenai.org. We hope our new dataset and ground truth annotations will provide an avenue for researchers to develop their algorithms and make comparisons.

5 Results

We assess our proposed algorithm on our dataset and compare its performance to previous methods. We expect the system being evaluated to return, for each figure extracted, a bounding box for both the figure and it's caption as well as the identifier of that figure (e.g., "Figure 1" or "Table 3") and the page number that the figure resides on. Figures with identifiers that did not exist in the hand built labels, or with incorrect page numbers, are considered incorrect. Otherwise a figure is judged by comparing the bounding boxes returned against the ground truth using the overlap score criterion from (Everingham and others 2010). Boxes are scored based on the area of intersection of the ground truth bounding box and the output bounding box divided by the area of the union between them. If the overlap score exceeds 0.80, we consider the output box to be correct, otherwise it is marked as incorrect. Caption box regions are scored using the same

[7] http://labelme2.csail.mit.edu/Release3.0/index.php