criterion. We consider an extraction to be correct if both the caption and the figure bounding boxes are correct according to the above criterion. Following this setup we evaluate our system and compare to the work of (Praczyk and Nogueras-Iso 2013). The work by (Praczyk and Nogueras-Iso 2013) does not return figure identifiers so we used regular expression to extract identifiers based on the first word of the caption text.

We also evaluate pdfimages (Poppler 2014), a popular tool for extracting embedded images from PDFs. We ran this tool on our dataset and filtered out the extracted images that were smaller than a square inch. We evaluate the results in this case using a much more lenient scoring criterion. We mark an extraction as correct if its size is within 80% of an annotated figure region on the corresponding page and wrong otherwise. Results of our evaluation on figures and tables can be found in Table 1 and Table 2 respectively. The outputs and evaluation of both our algorithm and the system by (Praczyk and Nogueras-Iso 2013) are available on our project website.

Our algorithm obtained around 96% precision at 92% recall for tables and figures. Despite being leniently scored, pdfimages performed extremely poorly. It is capable of getting correct results if a figure is encoded as a single embedded image within a document, but this is so rare in computer science papers, it was unable to even get 15% recall. The algorithm by (Praczyk and Nogueras-Iso 2013), although achieving high results on the domain of high energy physics (HEP), did not generalize well to the domain of computer science papers, achieving much lower recall and precision. Errors from (Praczyk and Nogueras-Iso 2013) are often due to mishandling cases where figures are close together, or allowing body text to be grouped into figure regions. In particular for papers from ICML, figure regions often included body text from the opposite column. Mistakes detecting captions were also a significant source of error. This indicates some of the heuristics used by (Praczyk and Nogueras-Iso 2013) failed to generalize from the HEP domain to computer science papers.

We also analyzed the sources of errors of our approach. Approximately a half of the errors produced by our algorithm were caused by non-standard formatting, such as, non-standard caption titles (e.g., Using "Figures 1 and 2" instead of "Figure 1:" and "Figure 2:"), or using small fonts for captions, or PDFs containing large numbers of extraneous operators that were detected by the text extraction tool (Poppler 2014) but were not visible on the document. The remaining half were caused by various text blocking errors, misclassifying blocks of text, or being unable to split regions containing multiple figures correctly.

Our algorithm takes less than two seconds to process a paper[8], and is therefore scalable to large datasets. In fact, we have run our method on 21,000 documents from the ACL corpus (Bird and others 2008) . The results are available on our project website.

6 Discussion

In this paper, we presented a novel approach (pdffigures) for extracting figures and tables from computer science papers. The contributions of our work include a method of identifying captions by exploiting the fact that captions are consistently formatted in scholarly documents, heuristics that are effective at separating body text and image text, and the insight that we can identify figures using the 'negative space' within body text as well the graphical elements within the PDF. We additionally released a novel dataset of computer science papers for figure extraction along with their ground truth labels. For future work, the figure assignment algorithm could be improved by using a more sophisticated scoring method or considering a wider diversity of region proposals for each caption, which could be used to provide resilience to errors in the previous steps. Finally, more carefully accounting for graphical elements, possibly integrating the kinds of clustering techniques that were used by (Praczyk and Nogueras-Iso 2013), might provide an avenue to improve results.

[8]On a single thread on a Macintosh OS X 10.9 with a 2.5GHz Intel core i7.

7 Acknowledgments

We thank Isaac Cowhey for annotating our dataset. We would also like to thank Oren Etzioni, Peter Clark, Isaac Cowhey, and Sam Skjonsberg for their helpful comments and reviews.

References

Aziz, et al. 2011. False-name manipulations in weighted voting games. *Journal of Artificial Intelligence Research*.

Bird, et al. 2008. The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *LREC*.

Chan, S. H., and Airoldi, E. M. 2014. A consistent histogram estimator for exchangeable graph models. *arXiv preprint arXiv:1402.1888*.

Choudhury, et al. 2013. Figure metadata extraction from digital documents. In *ICDAR*.

Everingham, et al. 2010. The pascal visual object classes (VOC) challenge. *International Journal of Computer Vision*.

Khusro, S.; Latif, A.; and Ullah, I. 2014. On methods and tools of table detection, extraction and annotation in pdf documents. *Journal of Information Science*.

Liu, W.; He, J.; and Chang, S.-F. 2010. Large graph construction for scalable semi-supervised learning. In *ICML*.

Lopez, et al. 2011. An automatic system for extracting figures and captions in biomedical pdf documents. In *BIBM*.

Neyshabur, et al. 2013. The power of asymmetry in binary hashing. In *NIPS*.

Poppler. 2014. Poppler. http://poppler.freedesktop.org/. Accessed: 2014-09-24.

Praczyk, P. A., and Nogueras-Iso, J. 2013. Automatic extraction of figures from scientific publications in high-energy physics. *Information Technology and Libraries*.

Smith, R. 2007. An overview of the tesseract ocr engine. In *ICDAR*.

Wu, et al. 2014. Citeseerx: AI in a digital library search engine.

Xu, S.; McCusker, J.; and Krauthammer, M. 2008. Yale image finder (YIF): a new search engine for retrieving biomedical images. In *Bioinformatics*. Oxford Univ Press.

Zanibbi, R.; Blostein, D.; and Cordy, J. R. 2004. A survey of table recognition. In *ICDAR*.