Table 3: Image Analysis Performance

| operation | accuracy |
|---|---|
| image segmentation | 98.12% |
| segment categorization | 100% |
| label identification | 96.97% |

| (label, structure) pair extraction | | | |
|---|---|---|---|
| method | precision | recall | F score |
| overall | 89.04% | 93.88% | 91.40% |
| easy | 97.69% | 100% | 98.83% |
| difficult | 77.55% | 93.65% | 84.84% |

Table 4: Text Analysis Performance

| method | precision | recall | F score |
|---|---|---|---|
| exact | 47.56% | 41.55% | 44.35% |
| rule-based | 28.66% | 94.12% | 43.95% |
| caCRF | 90.91% | 82.19% | 86.33 % |
| CRF | 90.91% | 82.19% | 86.33 % |

labels extracted from images, we do strict string matching to extract all the label appearances from text. The low precision is because the label text can appear in many scenarios other than indicating a chemical entity. Extracting irrelevant appearance harms precision. Moreover, a label can be mentioned in text with a slightly different format as introduced in images. For example, the label “(IIX)” can be referred to as “IIX”, “(IIX);”, “IIX,” etc. For this reason, strict string matching will miss many label appearances and has low recall. In the rule-based method, we specify rules about the composition of a label, similar to what we did in image analysis. As can be expected, this method has high recall, but generates many false positive and has low precision. The caCRF method achieves reasonable extraction performance. Moreover, the scheme of pre context selection significantly reduces the amount of data to be processed by CRF without influencing extraction accuracy. The amount of reduction is measured in terms of the number of tokens to be labeled by CRF, and we achieve 66.81% of reduction.

## Conclusion

In this work, we propose an IE scheme that explores the structural and language characteristics of chemical documents to bridge the gap between the visual content represented by images and the textual content represented by words. The scheme jointly mines the two media and is able to discover the knowledge which is otherwise lost by traditional single-media based mining systems.

## References

Banville, D. 2006. Mining chemical structural information from the drug literature. *Drug Discovery Today* 11(1-2):35–42.

Brecher, J. 1999. Name=struct: A practical approach to the sorry state of real-life chemical nomenclature. *Journal of Chemical Information and Computer Science* 39(6):943–950.

Chen, Y.; Spangler, S.; Kreulen, J.; and etc. 2009. Simple: A strategic information mining playform for ip excellence. *IBM Research Report*.

Cohen, W. W.; Wang, R.; and Murphy, R. F. 2003. Understanding captions in biomedical publications. In *ACM SIGKDD international conference on Knowledge discovery and data mining*, 499–504.

Deserno, T. M.; Antani, S.; and Long, L. R. 2009. Content-based image retrieval for scientific literature access. *Methods of Information in Medicine* 48(4):371–80.

Filippov, I. V., and Nicklaus, M. C. 2009. Optical structure recognition software to recover chemical information: Osra, an open source solution. *Journal of Chemical Information and Modeling* 49(3):740–743.

Futrelle, R. P. 2004. Handling figures in document summarization. In *Text Summarization Branches Out Workshop, 42nd Annual Meeting of the Association for Computational Linguistics*, 61–65.

Hamon, T., and Grabar, N. 2010. Linguistic approach for identification of medication names and related information in clinical narratives. *Journal of the American Medical Informatics Association* 17(5):549–554.

Hua, G., and Tian, Q. 2009. What can visual content analysis do for text based image search? In *IEEE international conference on Multimedia and Expo*, 1480–1483.

Klinger, R.; Kolářik, C.; Fluck, J.; Hofmann-Apitius, M.; and Friedrich, C. M. 2008. Detection of iupac and iupac-like chemical names. *Bioinformatics* 24:i268–i276.

Krauthammer, M.; Rzhetsky, A.; Morozov, P.; and Friedman, C. 2000. Using BLAST for identifying gene and protein names in journal articles. *Gene* 259(1-2):245–252.

Lafferty, J. D.; McCallum, A.; and Pereira, F. C. N. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning*, 282–289.

Lu, X.; Wang, J. Z.; Mitra, P.; and Giles, C. L. 2007. Deriving knowledge from figures for digital libraries. In *International conference on World Wide Web*, 1229–1230.

Park, J.; Rosania, G. R.; Shedden, K. A.; Nguyen, M.; Lyu, N.; and Saitou, K. 2009. Automated extraction of chemical structure information from digital raster images. *Chemistry Central journal* 3(1).

Sun, B.; Tan, Q.; Mitra, P.; and Giles, C. L. 2007. Extraction and search of chemical formulae in text documents on the web. In *International conference on World Wide Web*, 251–260.

Wagner, R. A., and Fischer, M. J. 1974. The String-to-String Correction Problem. *Journal of the ACM* 21(1):168–173.

Yan, R., and Naphade, M. 2005. Multi-modal video concept extraction using co-training. *Multimedia and Expo, IEEE International Conference on*.

Zimmermann, M., and Hofmann-Apitius, M. 2007. Automated extraction of chemical information from chemical structure depictions. *Drug Discovery* 12–15.