Optical Structure Recognition Application entry to CLEF-IP 2012

Igor V. Filippov[1], Dmitry Katsubo[2], and Marc C. Nicklaus[3]

[1] Chemical Biology Laboratory, SAIC-Frederick, Inc., Frederick National Lab, Frederick, Maryland, 21702, USA
igor.filippov@nih.gov,
WWW home page: http://cactus.nci.nih.gov/osra
[2] Life Sciences Department, European Patent Office, The Hague, The Netherlands
dmitry.katsubo@gmail.com
[3] Chemical Biology Laboratory, NCI, NIH, DHHS, Frederick National Lab, Frederick, Maryland, 21702, USA
mn1@helix.nih.gov

Abstract. We present our entry to CLEF 2012 Chemical Structure Recognition task. Our submission includes runs for both bounding box extraction and molecule structure recognition tasks using Optical Structure Recognition Application. OSRA is an open source utility to convert images of chemical structures to connection tables into established computerized molecular formats. It has been under constant development since 2007.

Keywords: image recognition, document analysis, chemoinformatics

1 Page segmentation

The general work-flow of OSRA has been presented at several meetings and conferences before: [1] [2] [3] [4]. A few modifications were made in the recently released version of OSRA to allow for more accurate bounding box coordinates reporting. Internally OSRA does not use bounding box paradigm, it relies instead on the minimum pairwise distance between points of different components. This allows to split (or keep together) objects which cannot be separated within a bounding box approach - i.e. imagine a larger molecule almost surrounding a smaller one. For this task we have submitted two runs - the first one was using tiffsplit to split multi-page TIFF images into separate pages, the second was using built-in OSRA facilities for page splitting. Surprisingly this lead to significantly different results which we can only attribute to the internal conversion of TIFF format going on within tiffsplit procedure.

Table 1 shows the result of page segmentation task when tiffsplit was used. The number of structures in the ground truth set was 5421, the total number of returned records was 8800. Tolerance shows the allowed margin of error in bounding box detection in pixels.

Tolerance	Precision	Recall	F1
0	0.43330	0.70338	0.53625
10	0.48989	0.79524	0.60629
20	0.50682	0.82273	0.62724
40	0.53580	0.86977	0.66310
55	0.54898	0.89116	0.67942

Table 1. OSRA page segmentation results using tiffsplit for page splitting

Tolerance	Precision	Recall	F1
0	0.70803	0.68622	0.69696
10	0.79311	0.76868	0.78070
20	0.82071	0.79543	0.80787
40	0.86696	0.84025	0.85340
55	0.88694	0.85962	0.87307

Table 2. OSRA page segmentation results using native page splitting

For the run using the OSRA native TIFF reading capabilities (Table 2) the number of returned records was 5254 and the precision overall was much higher. Both runs demonstrate competitive recall values.

2 Structure recognition

For structure recognition task the test set was split in two parts: the first one allowed for automatic result evaluation by using InChI keys in the same way as was applied at TREC-CHEM 2011 meeting. The second part was only possible to evaluate manually due to the presence of Markush-style atomic labels. The results are presented in Table 3.

Set	Structures	Recalled	%
Automatic	865	761	88%
Manual	95	38	40%
Total	960	799	83%

Table 3. Structure recognition task results

The results are consistent with those presented at the TREC-CHEM 2011 meeting where OSRA achieved second top-ranking score out of 6 participating projects.

Funding Disclaimer: This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under contract N01-CO-12400. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services,

nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government. This Research was supported in part by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research.

References

1. I. V. Filippov and M. C. Nicklaus. Optical Structure Recognition Software To Recover Chemical Information: OSRA, An Open Source Solution. *Journal of Chemical Information and Modeling*, 49(3):740–743, MAR 2009.
2. I. V. Filippov and M. C. Nicklaus Extracting chemical structure information: Optical structure recognition application. In *Proceedings of the Eight IAPR International Workshop on Graphics Recognition*, pages 133–142, 2009.
3. I. V. Filippov and M. C. Nicklaus and John Kinney Improvements in Optical Structure Recognition Application. *International Workshop on Document Analysis Systems (DAS 2010)*, 2010.
4. I. V. Filippov and Dmitry Katsubo and M. C. Nicklaus Optical Structure Recognition Application entry in Image2Structure task. In *Proceedings of the Twentieth Text REtrieval Conference (TREC 2011)*, 2011.