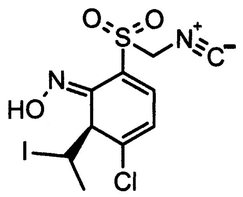
Performance of MolRec at TREC 2011
Overview and Analysis of Results

Noureddin M. Sadawi, Alan P. Sexton, and Volker Sorge
School of Computer Science, University of Birmingham

Abstract

- Image2Structure task in Chem track of TREC 2011



- We present an approach that models the principal recognition steps for molecule diagrams in a strictly rule based system, providing rules to identify the main components — atoms and bonds — as well as to resolve possible ambiguities.
- Initial vectorisation and OCR produces sets of geometric objects.
- Rule based system with fuzzy and strict parameters transforms this into a graph of bonds and atoms.
- Final graph can be output to MOL format files.
- Result on Trec was 95% and 94.9% correct recognition on two runs of 1000 images.

Vectorisation

- Connected component analysis, OCR and circle identification
- Skeletonisation and multiway junction splitting
- Douglas-Peucker73 for line simplification, then polyline separation

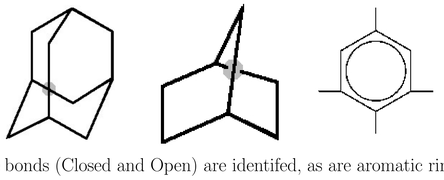
- Solid triangle and bold line identification using walking ball.

Rules to Capture Planar Bonds

- Planar Bonds can be: single, double or triple bonds

- Implicit nodes have C atoms where short lines end.

Rules for Bridge Bonds & Aromatic Rings



- Both types of bridge bonds (Closed and Open) are identifed, as are aromatic rings.

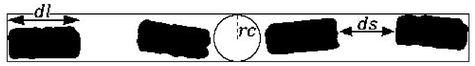
Rules to Capture Stereo Bonds

- "Up" Bonds: Solid wedge, Hollow wedge, Bold line.

- "Down" Bonds: Dashed wedge, Dashed bold line, Dashed normal line.

- Wavy bond

Example Rule: Capturing Dashed Bond



A dashed bond is captured using the following conditions:

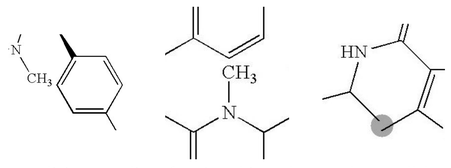
1. $L = \{l_1, \ldots, l_n\}$, where $n \geq 3$, is a set of line segments,
2. L is approximately colinear
3. Every element of L has length of approximately dl (dash length).
4. No two elements of L have a separation distance of less than the minimum of ds (dash separation).
5. Two elements of L, called the *end elements*, dash-neighbour wrt. ds precisely one other element of L. All other elements of L, called *internal elements*, dash-neighbour wrt. ds precisely two other elements of L.

Consequence L forms a dashed bond with endpoints given by the endpoints of the minimal line segment that contains $l_1, \ldots, l_n$.

- Full details and precise definitions to appear in [DRR 2012].

Things to Notice

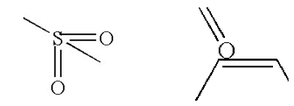
- Specific rules used to identify correct character grouping during vectorisation.
- There is a difference between Closed and Open nodes.



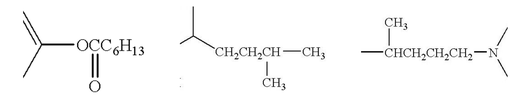
- If an open node has no explicit atom, lines forming this node must be parallel.
- Lines connected to explicit atoms must be pointing towards them.
- The rule based approach can be easily extended to accommodate more bond types.
- The strictness of the rule based approach makes it robust and resistant to misrecognising patterns that have similar appearance to some bond types.

Current Challenges

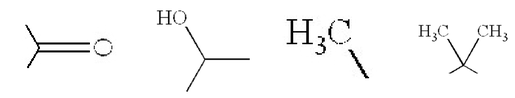
- Touching Components.



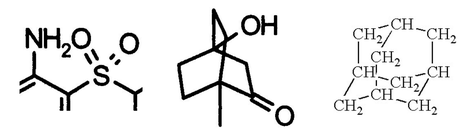
- Connectivity of Superatoms: it is unclear how to determine the appropriate permutation of connection possibilities between the superatom's substructure and the actual chemical structures in question.



- Broken Components, such as broken characters or lines.



- Some images can be very ambiguous, even for human readers.



- It is unclear how to determine the stereo-centre of dashed, bold, dashed bold bonds. Our current heuristics (based on graph connectivity analysis) resulted in 94% correct stereo-centre identification on the 168 such bonds in the TREC 2011 test set.

Results and Lessons Learned

- MolRec achieved a 95% and a 94.9% correct recovery rate on the two runs of 1000 images in the TREC11 test set.
- A total of 55 diagrams mis-recognised in one or other of the two runs with 61 reasons for diagram recognition failures in total.

Reason	#Images
Incorrect stereochemistry	10
Solid Circles without 3D Hydrogen Bond	5
Image has touching components	6
Image has broken characters	3
Incorrect character grouping	5
Connectivity of superatoms	3
Problematic Bridge Bonds	3
Unhandled bond type	1
Unrecognised syntax	5
Dashed wedge bonds mis-identified	15
Diagram caption confusion	5

Line finding and simplification: Hough transform approach disappointing. Our approach using Douglas-Peucker has proved to be fast, robust and particularly successful.

Solid triangle and bold line detection: Using our "walking ball" approach to find and orient solid triangles and bold lines has also proven to be very fast and robust.

Mining superatoms in MOL files: Lack of comprehensive superatom dictionaries led us to mine collections of MOL files for their superatom content. This has significantly increased the number of diagrams we can recognise automatically, although the superatom connectivity problem remains open.

Breaking joints: Explicitly breaking all connected joints allowed us to treat all bonds a uniform way, simplifying our code and the connection logic. This made dealing with implicit nodes much easier.