Text-based and Image-based Recognition and Extraction of Molecular Information from Figures and Figure Captions

Jungkap Park, Gus R. Rosania & Kazuhiro Saitou

University of Michigan, Ann Arbor
Outline

- Overview of Image-based Annotation
- ChemReader
- Annotation Strategy and Test Result
- Chemical Literature Database
- Preliminary Statistics
- Future Works
Why ChemReader?

Chemical Database
- PubChem
- ChemBank
- ChemDB
- ChemMine
- DrugBank
- GLIDA
- QueryChem
- ...

Scientific literature
- Journals
- Patents
- Books
- Papers
- Project reports
- Websites
- Theses
- ...

ChemReader
Searching for chemical information

- **The problems:**
  - Too many synonyms
  - Often referenced by chemical structure diagrams

Ex) Aspirin

- Acetylsalicylic acid (ASA)
- 2-acetyloxybenzoic acid
- acetylsalicylate
- Acylpyrin
- Colfarit
- Ecotrin
- Enterosarein
- Acenterine
- Polopiryna
- ........

[Chemical structure images]

P Vishweshwar et al, J. Am. Chem. 2005
The problems

• Need to identify related compounds

Aspirin


Advil
Image Based Annotation

- **Chemical database annotation using Chemical OCR**

  - **Chemical OCR system**
    - Extract 2D chemical structure diagram from literature
    - Convert to a standard chemical file format
    - CLIPE, ChemOCR, OSRA and **ChemReader**
Test Result

- **Recognition Test**

- **Annotation Test**
  - Tunable annotation strategy: Two different conditions for screening output structures

<table>
<thead>
<tr>
<th></th>
<th>Avg. Recall</th>
<th>Avg. Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test I</td>
<td>0.69</td>
<td>0.8</td>
</tr>
<tr>
<td>Test II</td>
<td>0.8</td>
<td>0.88</td>
</tr>
</tbody>
</table>
Ensemble Approach

- **Motivation**
  - Maximize the chance of including correct structure information by combining strengths of multiple chemical OCR systems

- **Rationale**
  - Different machine-vision algorithms could have different strengths in particular types of structures

Number of successful outputs produced by ChemReader or OSRA grouped by journal index.
Ensemble Approach

- **Use of multiple chemical OCR tools**

  - Two output structures for the same input structure become members of the ensemble
  - The ensemble approach enables to maximize chance of linking relevant entries in the annotation task
Annotation Test by Ensemble Approach

- **Result**
  - Total number of TP, FP and FN links

<table>
<thead>
<tr>
<th></th>
<th>TP</th>
<th>FP</th>
<th>FN</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChemReader</td>
<td>24592</td>
<td>30844</td>
<td>47631</td>
</tr>
<tr>
<td>OSRA</td>
<td>33105</td>
<td>21067</td>
<td>54995</td>
</tr>
<tr>
<td>Ensemble</td>
<td>45707</td>
<td>51535</td>
<td>55984</td>
</tr>
</tbody>
</table>

- Averaged recall and precision rates

<table>
<thead>
<tr>
<th></th>
<th>Avg. Precision</th>
<th>Avg. Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>ChemReader</td>
<td>0.563</td>
<td>0.569</td>
</tr>
<tr>
<td>OSRA</td>
<td>0.491</td>
<td>0.568</td>
</tr>
<tr>
<td>Ensemble</td>
<td>0.544</td>
<td>0.619</td>
</tr>
</tbody>
</table>
The need of image-based annotation

Motivation of Image-based annotation

- Many molecules are referenced by 2D structure diagrams in chemical literature due to the lack of standard names
- Image-based mining can uncover knowledge on such molecules that is otherwise inaccessible in chemical databases

How to validate?

- How chemical entities are referred in research articles?
- Comparison of text-based annotation and Image-based annotation
CAS Database
- The largest and commercially accessible chemical database
- Links to cited references (journals or patents) dating back to the beginning of the late 19th century

Sample set
- Keywords search: “Diabetes” and “small molecule”
  - 822 Journal articles
- Select 399 articles containing molecules being cited only once
- Download PDF files from publishers’ website
  - Total 346 full-text articles in PDF format
Extraction of chemical info from figures

- All figures and captions are extracted from articles

- **Image extraction**
  - Export images without modification of color depth, size or resolution
  - Snapshot tool only for vector graphics
  - Separation of chemical structure images

- **Chemical structure extraction**
  - 2D Chemical structure diagram from image files
  - Chemical names from caption text
  - Extracted chemicals are indexed by CAS Registry numbers (or InChI strings)
Construction of chemical literature database

- Extracted data is stored in a relational database as traceable assertions

```plaintext
CAS Database

- Article: 346
- Figure: 2129
- Non-chemical Image: 1082
- Chemical Diagram: 1679 + α
- Chemical Structure: 1873 + γ
- Caption: 3505 + β

* Red numbers denote the number of records in the database
```
Preliminary statistics on current data

- Identifying chemical diagrams or chemical names on progress

<table>
<thead>
<tr>
<th>Total number of linked molecules</th>
<th>cited in captions</th>
<th>cited in diagram</th>
<th>cited in both</th>
</tr>
</thead>
<tbody>
<tr>
<td>657 + α</td>
<td>1326 + β</td>
<td>110 + γ</td>
<td></td>
</tr>
</tbody>
</table>

- Over 278 molecules cited in chemical diagrams are missed by CAS
Text-based annotation using OSCAR3

- OSCAR3
  - Chemical documents processing tool (Corbett and Murray-Rust, 2008)
  - Identify chemical names, ontology terms and chemical data

- Chemical names in caption text
  - Number of captions tested: 334
  - Number of chemical names = 1087
  - Number of chemical names extracted by OSCAR = 1814
  - Number of correctly identified = 806
  - Precision = 0.444
  - Recall = 0.741
What we can do with the database

- **Statistical Analysis**
  - How molecules are cited first? By diagrams or names?
  - How many molecules are cited only by diagrams?
  - How many molecules are not indexed by CAS?

2D Chemical diagrams in articles are important data objects for mining chemical literature
Validation of Image-Based Annotation

ChemReader is effective?

- Chemical structures cited only by diagrams and missed by CAS
- Chemical structures incorrectly annotated by text-based approach

Image-based approach can uncover knowledge that are inaccessible otherwise
Integration of Image-based and Text-based Multi-modal extraction from chemical literature
- Text-based mining enables to extract textual descriptors as well as chemical names
- Graphical Mining
- Uncover the contextual scientific knowledge

Ensemble approach
- Strengths of image-based and text-based techniques
- Increase annotation accuracy
Conclusion

- Significant fraction of molecules is referenced by chemical diagrams only, and a chemical OCR system can be effective in annotating articles with these molecules.

- Constructed database will facilitate research in chemical literature mining for the design, training and testing of algorithms for chemical structure extraction and chemical database annotation.
Acknowledgement

- Polyergic Informatics, LLC
- Small Company Innovation Program, College of Engineering
- Michael Conlin
- Ye Li
- Christof Smith
- Caroline Yee
- Bethany Harris
Thank you!