Mining for Chemistry in Text and Images. A Real-World Example revealing the Challenge, Scope, Limitation and Usability of the current Technology

InfoChem GmbH, Landsberger Strasse 408, Munich, 81241, Germany
### Paracetamol Connection Table

<table>
<thead>
<tr>
<th>CAS Nr.: 103-90-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>11 11 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>1.4944 -0.4750 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>1.4932 -1.3023 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>2.2081 -1.7152 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>2.9245 -1.3019 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>2.9216 -0.4713 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>2.2063 -0.0622 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>0.7784 -1.7143 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>3.6346 -0.0561 0.0000 N 0 0 0 0 0 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>4.3458 -0.4667 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>4.3459 -1.2917 0.0000 O 0 0 0 0 0 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>5.0602 -0.0541 0.0000 C 0 0 0 0 0 0 0 0 0 0 0 0</td>
</tr>
<tr>
<td>5 6 2 0 0 0 0</td>
</tr>
<tr>
<td>6 1 1 0 0 0 0</td>
</tr>
<tr>
<td>1 2 2 0 0 0 0</td>
</tr>
<tr>
<td>2 7 1 0 0 0 0</td>
</tr>
<tr>
<td>3 4 2 0 0 0 0</td>
</tr>
<tr>
<td>5 8 1 0 0 0 0</td>
</tr>
<tr>
<td>8 9 1 0 0 0 0</td>
</tr>
<tr>
<td>4 5 1 0 0 0 0</td>
</tr>
<tr>
<td>9 10 2 0 0 0 0</td>
</tr>
<tr>
<td>2 3 1 0 0 0 0</td>
</tr>
<tr>
<td>9 11 1 0 0 0 0</td>
</tr>
<tr>
<td>M END</td>
</tr>
</tbody>
</table>
Conventional Database Building: Concept

### Advantages
- High quality
- Currently no alternative

### Disadvantages
- Time-consuming, Time-gap
- Not-comprehensive patent extraction
- Legacy data extraction not affordable
- High costs:
  - 5-7 USD per reaction (depending on source)
  - 2,5-4 USD per structure (depending on source)

---

**Print Version**

**Structure / Reaction Database Building**
- Structures/RXN
- Factual Data
- Catalysts
- Solvents

**Connection Tables**

**Manual Abstraction**
Mining for Chemistry in Text and Images

- Chemical Named Entity Recognition (NER):
  - Automatic identification of chemical names in text documents
  - Generation of connection table

- Image Recognition (IR):
  - Images of chemical structures in text are not computer searchable
  - IR enables generation of computer readable structures

- Mining for Chemistry goal:
  - Automatic generation of structure and reaction databases
Activities in the Area of Chemical Mining

- Chemical Named Entity Recognition:
  - Oscar (Corbett, Murray-Rust, Teufel et al.)
  - Chemical Entity Relationships Skill Cartridge for Luxid (Temis)
  - SureChem (former ReelTwo)
  - ChemBrowser (IBM / InfoChem)
  - ICANNOTATOR (InfoChem)
  - …

- Image Recognition:
  - chemoCR (Fraunhofer SCAI / InfoChem)
  - CLiDE Pro (Keymodule Ltd.)
  - OSRA (NIH)
  - …

- Pilot project InfoChem / SCAI:
  - Status of some of the technologies used in these areas
Our Real-World Example

Pilot project with highly heterogeneous source set:

- Sample set: 15 titles (journals, book series)
  - contains chemical compounds in text and images
  - covers different areas of chemistry
  - diverse source formats (scanned, text, pdfs, etc.)

- Aim of project: Proof of concept on sample set
  Assessment of results and quality
Overview of Approach and Applied Technology

Manual abstraction of chemical names

NER \rightarrow N2S

Database

Automatic chemical verification

Comparison (quantitative)

"Similarity MCD"

Pdf to Text

NER

ICANNOTATOR

Page segmentation \rightarrow Image classifier \rightarrow Chemical recognition

chemoCR (Fraunhofer)

Manual abstraction of structures from images
Selection of Sample Set

• Sources: 15 titles covering time period 2000-2008
  • 10 journals
  • 5 book series

• Definition of random subset:
  • 10 articles of each title equally distributed over the time period
  • 3 pages equally distributed over every article are processed

• Total: 450 pages containing text and images
Challenges NER: Scanned Documents

- Visible and extracted text not identical

The structure of the only representatives of unsymmetrical 3,4'- and 4,5-disubstituted 2,2'-bithiophenes 3,4'-dibromo-2,2'-bithiophene (17) [29] and 4-(2-thienyl)-5-phenyl-2,2'-bithiophene (18) [30] was determined. In the molecule of the first compound slight deviation from planarity is observed, and as a result the torsion angle $S_{11} - C_{12} - C_{13} - S_{14}$ is $175.0^\circ$. The bond lengths and the angles of the heterocycle are comparable with the values determined for the other bithiophenes. The heterocycles in the 2,2'-bithiophene fragment of compound 18 are

**OCR results:**

The structure of the only representatives of unsymmetrical 3,4'- and 4,5-disubstituted 2,2'-bithiophenes 3,4'-dibromo-2,2'-bithiophene (17) [29] and 4-(2-thienyl)-5-phenyl-2,2'-bithiophene (18) [30] was determined. In the molecule of the first compound slight deviation from planarity is observed, and as a result the torsion angle $S_{11} - C_{12} - C_{13} - S_{14}$ is $175.0^\circ$. The bond lengths and the angles of the heterocycle are comparable with the values determined for the other bithiophenes. The heterocycles in the 2,2'-bithiophene fragment of compound 18 are
Challenges NER: „Native“ pdf

• Visible and extracted text not identical

1. Greek letters and apostrophes:
   
   \[7\alpha,8\alpha,3',4']-N'-(Phenyl)succinimido-6,14-endo-etheno-6,7,8,14-tetrahydrothebaine (VI) was synthesized using the method described in [16];

   \textit{PdfToText results:}

   \[7a,8a,3\ddagger,4\ddagger]-N\ddagger-(Phenyl)succinimido-6,14-endo-etheno….

2. Notes, indexes:

   a4-Azochromotropic acid pentylfuorone

   \textit{PdfToText results:}

   a4-Azochromotropic acid pentylfuorone
More Challenges NER

• Misspelled names (typing errors in original document)

• Errors introduced by N2S conversion tools

• Incomplete and erroneous dictionaries

• …
Challenges Image Recognition (1)

• Brackets are converted into bonds:
Challenges Image Recognition (2)

• Atom numbers in ring are recognized as part of the structure:
Challenges Image Recognition (3)

• Variable point of attachment bonds are not recognized:

```
R1
R2

+ H-N
R2
R3

Pd(OAc)2 (1)
IPr-HCl (5)
KO'Bu (3)
Toluene, 10

16_153_19_1
```

Created from 15_153_19_1
Verification and Results NER

Recall  Precision  F-score
79%        69%        73%
Verification Image Recognition: SimilarityMCD

• Rating method used to compare manually and automatically abstracted chemical structures
• Basic idea is to rate the chemical difference of the two compared molecules

\[
\text{MCD} = \text{breaking a minimum number of bonds in the reactant and making a minimum number of bonds in the product}
\]

• The identical parts and different parts of the two compared structures are rated in a percent value from 0% to 100% based on the Largest Common Subgraph (LCS) and the Minimum Chemical Distance (MCD)
Verification Image Recognition: Rating

• All bonds of the two molecules get different ratings for single, double and triple bonds, for C-C and C-hetero atoms bonds, depending on bond stability

\[
\text{SimilarityMCD}(\text{Mol1, Mol2}) = \frac{\text{Mapped Value}}{\text{Total Value}} \times 100 \ [\%] 
\]

“Mapped Value”

\[= \sum \text{ of ratings for mapped bonds and atoms}\]

“Total Value” (per definition 100% rating)

\[= \sum \text{ of all bond ratings of the two compared molecules}\]
Verification Image Recognition: Examples

1. 100%
2. 97.73%
3. 57.97%
Results Image Recognition (Source No. 5)

![Graph showing similarity percent values for processed structures]

- P11
- P4
- P7

Processed structures:
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39 41 43 45 47 49 51 53 55 57 59 61 63 65

Similarity/MCD percent values:
0 10 20 30 40 50 60 70 80 90 100
## Results Image Recognition (P11)

<table>
<thead>
<tr>
<th>Source</th>
<th>Overall Similarity MCD</th>
<th>All Sources 100% Match</th>
<th>Best Source 100% Match</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>72%</td>
<td>33%</td>
<td>58%</td>
</tr>
</tbody>
</table>

The chart shows the count of structures and their similarity MCD percent values across different sources. The data indicates that the overall similarity MCD is 72%, with the best source achieving a 100% match at 58%.
Conclusions

• Using a highly heterogeneous sample set from a real world example we have:
  • illustrated the challenges of chemical NER and image recognition
  • demonstrated with a quantitative comparison the quality of ICANNOTATOR
  • explained how the SimilarityMCD algorithm enables quantitative comparison and can be used to optimize chemoCR parameter sets

• We have achieved up to 93% recall for text and up to 58% exactly recognized structures

• All components are in the process of further development
Acknowledgements

- Prof. Dr. Martin Hofmann-Apitius, Fraunhofer SCAI
- Dr. Marc Zimmermann and his Team, Fraunhofer SCAI

InfoChem GmbH:

www.infochem.de
www.spresi.com
info@infochem.de