Analysing and Classifying Names of Chemical Compounds with CHEMorph

Stefanie Anstein  Gerhard Kremer

IMS, University of Stuttgart

April 11, 2006
Example Analysis

7-hydroxyheptan-2-one

compd(ane(7*C), pref([1*[7]-hydroxy]), suff([1*[2]-one]))

CC(=O)CCCCCO

ALCOHOL, KETONE, ...
Example Analysis

7-hydroxyheptan-2-one

compd(ane(7*C),pref([1*[7]-hydroxy]),suff([1*[2]-one]))

CC(=O)CCCCCO

ALCOHOL,KETONE,...
Example Analysis

\[
\begin{align*}
\text{H} & \quad \text{O} & \quad \text{H} & \quad \text{H} & \quad \text{H} & \quad \text{H} & \quad \text{H} & \quad \text{H} \\
\quad & \quad & \quad & \quad & \quad & \quad & \quad & \quad \\
\text{H} & \quad \text{C} & \quad \text{C} & \quad \text{C} & \quad \text{C} & \quad \text{C} & \quad \text{C} & \quad \text{C} & \quad \text{C} & \quad \text{O} & \quad \text{H} \\
\quad & \quad & \quad & \quad & \quad & \quad & \quad & \quad & \quad & \quad & \quad \\
\text{H} & \quad \text{H} & \quad \text{H} & \quad \text{H} & \quad \text{H} & \quad \text{H} & \quad \text{H} & \quad \text{H} & \quad \text{H} & \quad \\
\end{align*}
\]

7-hydroxyheptan-2-one

\[
\text{compd(ane(7*C),pref([1*[7]-hydroxy]),suff([1*[2]-one]))}
\]

\[
\text{CC(=O)CCCCCO} \quad \text{ALCOHOL,KETONE, ...}
\]
Example Analysis

7-hydroxyheptan-2-one

compd(ane(7*C),pref([1*[7]-hydroxy]),suff([1*[2]-one]))

CC(=O)CCCCCO  ALCOHOL,KETONE,...
Motivation & Background

- life sciences . . .
  and the amount of biomedical data

- terminology . . .
  and biochemical nomenclature
Motivation & Background

- life sciences . . .
  and the amount of biomedical data

- terminology . . .
  and biochemical nomenclature
Challenges

- term reference
- coreferences
Challenges

- term reference
- coreferences

R-0. 1.7.3 (IUPAC nomenclature of organic compounds):
Addition of the vowel “o”.
For euphonic reasons, the vowel “o” is sometimes inserted between consonants.
**Modules Overview**

- **name**
  - parser
    - semantic representation
      - SMILES string generator
      - classifier
        - classes
        - SMILES string
Modules Overview

- **name**
  - **parser**
    - **semantic representation**
      - **SMILES string generator**
        - **SMILES string**
      - **classifier**
        - **classes**
Modules Overview

- Name
  - Parser
    - Semantic representation
      - SMILES string generator
        - SMILES string
      - Classifier
        - Classes
## Name Types

<table>
<thead>
<tr>
<th>Name Type</th>
<th>Fully Specified</th>
<th>Underspecified</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Systematic</strong></td>
<td>7-hydroxyheptan-2-one</td>
<td>heptanone</td>
</tr>
<tr>
<td><strong>Trivial</strong></td>
<td>benzene</td>
<td>∅</td>
</tr>
<tr>
<td><strong>Semi-systematic</strong></td>
<td>benzene-1,3,5-triacetic acid</td>
<td>dihydrobenzene</td>
</tr>
<tr>
<td><strong>Class</strong></td>
<td>∅</td>
<td>alcohol</td>
</tr>
<tr>
<td><strong>Semi-systematic</strong></td>
<td>∅</td>
<td>2-deoxysugar</td>
</tr>
</tbody>
</table>
Parser

\[ \text{compd}( \text{ane}(7*C) , \text{pref}( [??*[7]-\text{hydroxy}] ) , \text{suff}( [??*[2]-\text{one}] ) ) \]
compd( ane(7*C) , pref( [??*[7]-hydroxy] ) ) ,
suff( [??*[2]-one] )

Stefanie Anstein, Gerhard Kremer
CHEMorph
Parser

Stefanie Anstein, Gerhard Kremer
Parser

\[
\text{compd} \left( \text{ane}(7\ast C) , \text{pref}( \text{[??*[7]-hydroxy]} ) , \text{suff}( \text{[??*[2]-one]} ) \right)
\]
**Parser**

```
7
| loc [7] | hyphen ⊘ | hydroxy | hept | an | - | 2 | - | one
| prefix [??*[7]-hydroxy] | parent nonsugar ane(7*'C') | multi 7 parent_suffix λ(X, ane(X*'C')) | hyphen ⊘ | loc [2] | hyphen ⊘ | suff one
|
| locant [??*[7]] |
| prefix [??*[7]-hydroxy] |
| parent_nonsugar ane(7*'C') |
| suffix [??*[2]-one] |

compd( ane(7*C), pref([??*[7]-hydroxy]), suff([??*[2]-one]) )
```
SMILES String Generator

- representation of single chain elements
- consistency check
- underspecification:
  
  underspecified( CC(=O)CCCCC , [1,3,4,5,6,7]-hydroxy)
SMILES String Generator

- representation of single chain elements
- consistency check
- underspecification:

```plaintext
underspecified( CC(=O)CCCCC , [{1,3,4,5,6,7}-hydroxy] )
```
SMILES String Generator

- representation of single chain elements
- consistency check
- underspecification:
  
  underspecified( CC(=O)CCCCC , [{1,3,4,5,6,7}-hydroxy] )
Classifier

<table>
<thead>
<tr>
<th>morpheme</th>
<th>class</th>
</tr>
</thead>
<tbody>
<tr>
<td>hydroxy-</td>
<td>ALCOHOL</td>
</tr>
<tr>
<td>-ol</td>
<td></td>
</tr>
<tr>
<td>cyclo-</td>
<td>CYCLOALKANE</td>
</tr>
<tr>
<td>&amp; -ane</td>
<td></td>
</tr>
</tbody>
</table>

\\( \text{compd( ane(7*C), pref([1*[7]-hydroxy]), suff([1*[2]-one]) )} \rightarrow \text{ALKANE, ALCOHOL, KETONE} \)

\\( \text{compd( ene(??[??],ane(4*’C’)), pref([]), suff([])) } \rightarrow \text{ALKENE} \)
**Classifier**

<table>
<thead>
<tr>
<th>morpheme</th>
<th>class</th>
</tr>
</thead>
<tbody>
<tr>
<td>hydroxy-</td>
<td>ALCOHOL</td>
</tr>
<tr>
<td>-ol</td>
<td></td>
</tr>
<tr>
<td>cyclo-</td>
<td>CYCLOALKANE</td>
</tr>
<tr>
<td>&amp; -ane</td>
<td></td>
</tr>
</tbody>
</table>

- compd( ane(7*C) , pref([1*[7]-hydroxy]) , suff([1*[2]-one]) )
  → ALKANE, ALCOHOL, KETONE

- compd( ene(??*[??],ane(4*'C')) , pref([]) , suff([]) )
  → ALKENE
Classifier

\[
\begin{array}{c|c}
\text{morpheme} & \text{class} \\
\hline
\text{hydroxy- | -ol} & \text{ALCOHOL} \\
\text{cyclo- & -ane} & \text{CYCLOALKANE} \\
\end{array}
\]

- \text{compd( \text{ane(7\text{*C})}, \text{pref([1*[7]-\text{hydroxy}]})}, \text{ suff([1*[2]-\text{one}]) } )
  \rightarrow \text{ALKANE, ALCOHOL, KETONE}

- \text{compd( \text{ene(??*[??],ane(4*’C’)}}, \text{ pref([])}}, \text{ suff([]) } )
  \rightarrow \text{ALKENE}
## Classifier

<table>
<thead>
<tr>
<th>morpheme</th>
<th>class</th>
</tr>
</thead>
<tbody>
<tr>
<td>hydroxy-</td>
<td>-ol</td>
</tr>
<tr>
<td>cyclo- &amp; -ane</td>
<td></td>
</tr>
</tbody>
</table>

- compd( `ane(7*C)` , pref([1*[7]-hydroxy]) , suff([1*[2]-one]) )
  →  ALKANE, ALCOHOL, KETONE

- compd( `ene(??*[??],ane(4*’C’))` , pref([]) , suff([]) )
  →  ALKENE
Results & Applications

- SMILES string and classification
- underspecification
- term reference
- coreference resolution
- database curation and ontology acquisition
Results & Applications

- SMILES string and classification
- underspecification
- term reference
- coreference resolution
- database curation and ontology acquisition
Results & Applications

- SMILES string and classification
- underspecification
- term reference
- coreference resolution
- database curation and ontology acquisition
Results & Applications

- SMILES string and classification
- underspecification
- term reference
- coreference resolution
- database curation and ontology acquisition
Results & Applications

- SMILES string and classification
- underspecification
- term reference
- coreference resolution
- database curation and ontology acquisition
Conclusion & Outlook

- feasible, extendable and transferable approach
- extend grammar and lexicon
- elaborate SMILES and classification
- sophisticated linguistic analysis $\rightarrow$ database curation
- term identification $\rightarrow$ text processing applications
Conclusion & Outlook

- feasible, extendable and transferable approach
- extend grammar and lexicon
- elaborate SMILES and classification
- sophisticated linguistic analysis → database curation
- term identification → text processing applications
Conclusion & Outlook

- feasible, extendable and transferable approach
- extend grammar and lexicon
- elaborate SMILES and classification
- sophisticated linguistic analysis $\rightarrow$ database curation
- term identification $\rightarrow$ text processing applications
Acknowledgements

Stefanie Anstein
Uwe Reyle
Jasmin Šarić
EML Research gGmbH
Schönen Dank.
IUPAC Nomenclatures

Amino Acids and Peptides
Biochemical thermodynamics
Branched nucleic acids
**Carbohydrates**
Carotenoids
Corrinoids (vitamin B12)
Cyclitols
Electron transport proteins
Enzyme kinetics
Enzyme nomenclature
EC 1 Oxidoreductases
EC 2 Transferases
EC 3 Hydrolases
EC 4 Lyases
EC 5 Isomerases
EC 6 Ligases
Folic acid
Glycolipids
Glycoproteins
myo-Inositol numbering
Lignan Nomenclature
Lipid Nomenclature
Multienzymes
Multiple forms of enzymes
Nucleic acid constituents
Nucleic acid sequence
Peptide hormones
Phosphorus containing compds
Polymerized amino acids
Polypeptide conformation
Polynucleotide conformation
Polysaccharide conformation
Prenol nomenclature
Pyridoxal (vitamin B6)
Quinones w. an Isoprenoid Chain
Retinoids
Steroids
Tetrapyrroles
Tocopherols (vitamin E)
Translation Factors
Vitamin D

Organic Chemistry
**COMPOUND: C03802**

<table>
<thead>
<tr>
<th>Entry</th>
<th>C03802</th>
<th>Compound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Ribonucleoside triphosphate</td>
<td></td>
</tr>
<tr>
<td>Formula</td>
<td>C5H12O13P3R</td>
<td></td>
</tr>
<tr>
<td>Mass</td>
<td>372.9489</td>
<td></td>
</tr>
<tr>
<td>Structure</td>
<td><img src="image" alt="Compound Structure" /></td>
<td></td>
</tr>
</tbody>
</table>

**Reaction**  R04315

**Enzyme**  1.17.4.2

**Other DBs**  PubChem: 6551

**LinkDB**  All DBs

**KCF data**  Show

KEGG: Kyoto Encyclopedia of Genes and Genomes
**COMPOUND: C03802**

<table>
<thead>
<tr>
<th>Entry</th>
<th>C03802</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Ribonucleoside triphosphate</td>
</tr>
<tr>
<td>Formula</td>
<td>C5H12O13P3R</td>
</tr>
<tr>
<td>Mass</td>
<td>372.9489</td>
</tr>
<tr>
<td>Structure</td>
<td>![Structure Image]</td>
</tr>
</tbody>
</table>

**Reaction**

| Reaction | R04315 |

**Enzyme**

| Enzyme | 1.17.4.2 |

**Other DBs**

- PubChem: 6551

**LinkDB**

- All DBs

**KCF data**

- Show

---

**COMPOUND: C00699**

<table>
<thead>
<tr>
<th>Entry</th>
<th>C00699</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>NTP</td>
</tr>
<tr>
<td>Formula</td>
<td>C5H12O13P3R</td>
</tr>
<tr>
<td>Mass</td>
<td>372.9489</td>
</tr>
<tr>
<td>Structure</td>
<td>![Structure Image]</td>
</tr>
</tbody>
</table>

---

**COMPOUND: C00201**

<table>
<thead>
<tr>
<th>Entry</th>
<th>C00201</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Nucleoside triphosphate</td>
</tr>
<tr>
<td>Formula</td>
<td>C5H12O13P3R</td>
</tr>
<tr>
<td>Mass</td>
<td>372.9489</td>
</tr>
<tr>
<td>Structure</td>
<td>![Structure Image]</td>
</tr>
</tbody>
</table>
Reaction: $3'$-phosphoadenylylsulfate + an alcohol = adenosine 3',5'-bisphosphate + an alkyl sulfate.

Other name(s): Hydroxysteroid sulfotransferase.

Comments: Primary and secondary alcohols, including aliphatic alcohols, ascorbate, chloramphenicol, ephedrine and hydroxysteroids, but not phenolic steroids, can act as acceptors (cf. Ec 2.8.2.15).
Ozonolysis of Alkenes and Study of Reactions of Polyfunctional Compounds: LXIII. A New Procedure for Direct Reduction of 1-Methylcycloalkene Ozonolysis Products to Hydroxyketones

Authors: Ishmuratov G.Y.¹; Kharisov R.Y.¹; Yakovleva M.P.¹; Botsman O.V.¹; Muslukhov R.R.¹; Tolstikov G.A.¹
Publisher: MAIK Nauka Interperiodica

Abstract:
A procedure was proposed for direct reduction of peroxide products resulting from ozonolysis of 1-methylcycloalkenes to the corresponding hydroxyketones by the action of sodium triacetoxyhydridoborate.

Language: English
Document Type: Regular paper
Affiliations: ¹: Institute of Organic Chemistry, Ufa Research Center, Russian Academy of Sciences, pr. Oktyabrya 71, Ufa, 450054 Bashkortostan, Russia
<table>
<thead>
<tr>
<th>Entry</th>
<th>C01801</th>
<th>Compound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name</td>
<td>Deoxyribose; 2-Deoxy-beta-D-erythro-pentose; Thyminose; 2-Deoxy-D-ribose</td>
<td></td>
</tr>
<tr>
<td>Formula</td>
<td>C5H10O4</td>
<td></td>
</tr>
<tr>
<td>Mass</td>
<td>134.0579</td>
<td></td>
</tr>
</tbody>
</table>