

Pyridines, Pyridine and Pyridine Rings: Disambiguating Chemical Named Entities

Peter Corbett

- Unilever Centre for Molecular Sciences Informatics
University of Cambridge, Chemical Laboratory

Colin Batchelor

- Royal Society of Chemistry

Ann Copestake

- Natural Language and Information Processing Group
University of Cambridge, Computer Laboratory

Background

- Chemical Named Entity Guidelines
- 5 NE classes
 - Dominant (~95%) class is CM (chemical)
- Inter-Annotator Agreement
 - $F = 93\%$
- Applied to corpus of 42 chemistry papers
 - Provided by Royal Society of Chemistry
 - Covers all chemical subdomains
 - Overlap with other domains, e.g. biochemistry, materials science, environmental science

Annotation of Chemical Named Entities

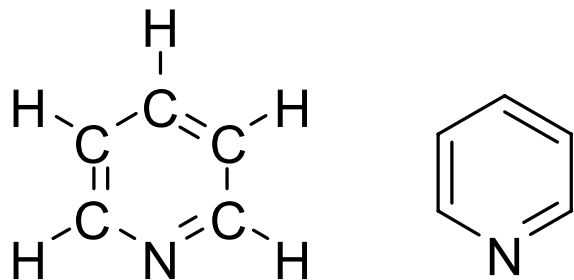
Peter Corbett, Colin Batchelor, Simone Teufel

Proceedings of BioNLP 2007, 57-64

A Problem

- CM does not distinguish between
 - Specific chemical compounds
 - Classes of chemical compounds
 - Parts of chemical compounds
- Early versions of guidelines attempted to deal with this, using simple name-internal cues (e.g. plural => class)
- Problem: Polysemy

Pyridine



“The green residue was dissolved in pyridine”

Properties

Molecular formula

C_5H_5N

Molar mass

79.101 g/mol

Appearance

colourless liquid

Density

0.9819 g/cm³,
liquid

Melting point

-41.6 °C

Boiling point

115.2 °C

Solubility in water

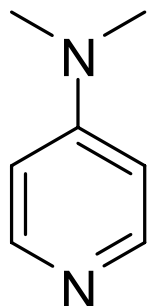
Miscible

Viscosity

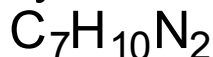
0.94 cP at 20 °C

(From Wikipedia)

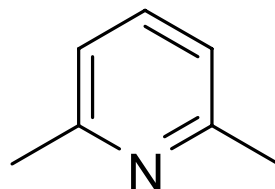
Pyridines



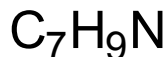
4-Dimethylaminopyridine



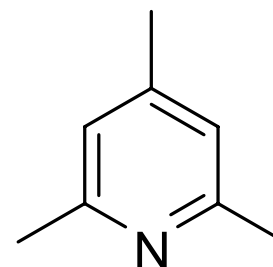
m.p. 110-113 °C



2,6-lutidine



m.p. -5.8 °C



2,4,5-collidine



m.p. -46 °C


“Typically this reaction may be carried out in the presence of a pyridine such as an alkyipyridine...”



Search ChEBI

Search

Help

- ▀ ChEBI Home
 - ▀ Advanced Search
 - ▀ Browse
 - ▀ Downloads
 - ▀ Documentation
 - ▀ Developer Resources
 - ▀  Preferences
 - ▀ Contact ChEBI
-
- ▀ Printer Friendly View

EBI > Databases > Small Molecules > ChEBI > Main

pyridines (CHEBI:26421)

Main

Automatic Xrefs

ChEBI Name	pyridines
ChEBI ID	CHEBI:26421
Last Modified	19 March 2008

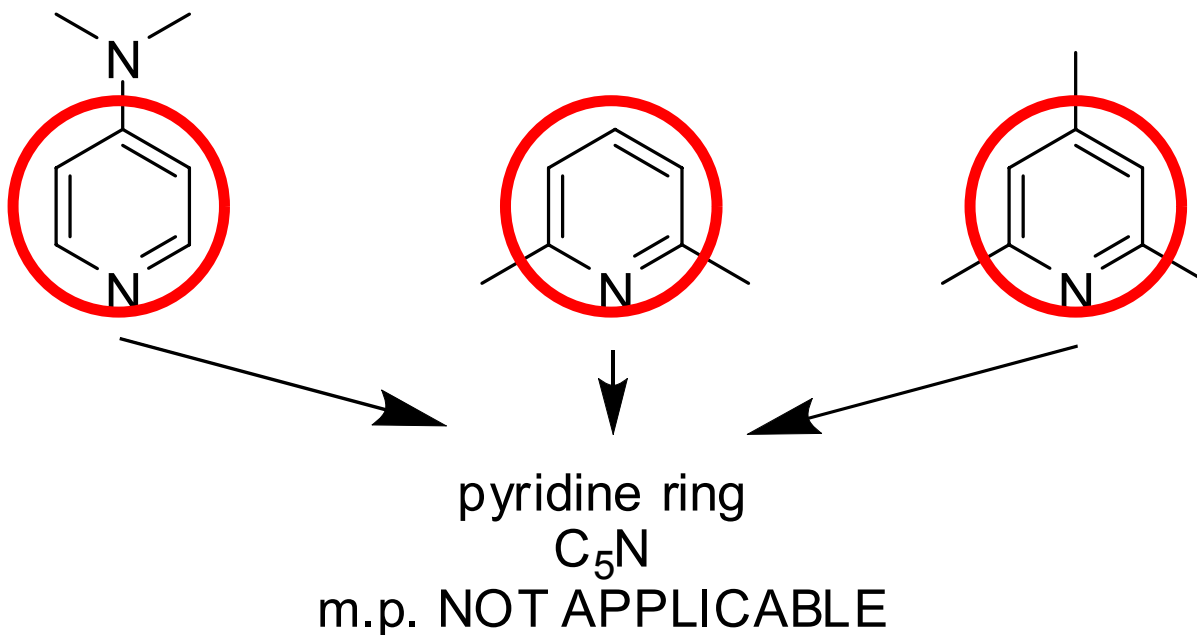
ChEBI Ontology

 Tree view

Parents	pyridines (CHEBI:26421) is a organic heteromonocyclic compounds (CHEBI:25693)
	pyridines (CHEBI:26421) is a organonitrogen heterocyclic compounds (CHEBI:38101)
	pyridinemonocarboxylic acids (CHEBI:26420) is a pyridines (CHEBI:26421)
	cyanopyridines (CHEBI:23438) is a pyridines (CHEBI:26421)
	hydroxypyridines (CHEBI:24745) is a pyridines (CHEBI:26421)
	methylpyridines (CHEBI:25340) is a pyridines (CHEBI:26421)
	pyridinecarboxamides (CHEBI:25529) is a pyridines (CHEBI:26421)
	pyridine (CHEBI:16227) is a pyridines (CHEBI:26421)
	pyridine alkaloids (CHEBI:26416) is a pyridines (CHEBI:26421)
	tetrahydropyridines (CHEBI:26921) is a pyridines (CHEBI:26421)
	vitamin B ₆ (CHEBI:27306) is a pyridines (CHEBI:26421)
	pyridinedicarboxylic acids (CHEBI:36112) is a pyridines (CHEBI:26421)
	sodium picosulfate (CHEBI:32147) is a pyridines (CHEBI:26421)
	<i>N</i> -glycosylpyridines (CHEBI:36979) is a pyridines (CHEBI:26421)
	aminoalkylpyridines (CHEBI:38198) is a pyridines (CHEBI:26421)
	aminoacylpyridines (CHEBI:38208) is a pyridines (CHEBI:26421)
	pyridinemonocarboxylates (CHEBI:38181) is a pyridines (CHEBI:26421)
	pyridones (CHEBI:38183) is a pyridines (CHEBI:26421)
	pyridinocarbonylhydrazides (CHEBI:38187) is a pyridines (CHEBI:26421)

pyriminemonocarboxylic acids ([CHEBI:20420](#)) is a pyrimidines ([CHEBI:20420](#))
cyanopyridines ([CHEBI:23438](#)) is a pyridines ([CHEBI:26421](#))
hydroxypyridines ([CHEBI:24745](#)) is a pyridines ([CHEBI:26421](#))
methylpyridines ([CHEBI:25340](#)) is a pyridines ([CHEBI:26421](#))
pyridinecarboxamides ([CHEBI:25529](#)) is a pyridines ([CHEBI:26421](#))
pyridine ([CHEBI:16227](#)) is a pyridines ([CHEBI:26421](#))
pyridine alkaloids ([CHEBI:26416](#)) is a pyridines ([CHEBI:26421](#))
tetrahydropyridines ([CHEBI:26921](#)) is a pyridines ([CHEBI:26421](#))
vitamin B₃ ([CHEBI:27306](#)) is a pyridines ([CHEBI:26421](#))

Pyridine Rings



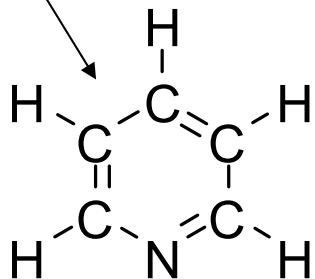
“In this paper, we report two pyridine-containing triphenylbenzene derivatives of 1,3,5-tri(m-pyrid-3-yl-phenyl)benzene...”

Pyridine is a pyridine

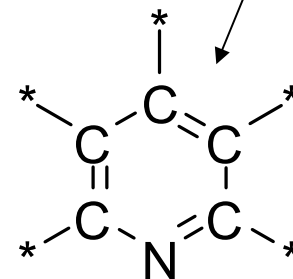
- One Sense Per Discourse does not apply
- Found using Google
 - “A pyridine such as pyridine”
 - “Pyridines such as pyridine itself”
 - “Pyridines including pyridine, 4-dimethylaminopyridine...”

Denotation

“The green residue
was dissolved in
pyridine”



“Typically this reaction may
be carried out in the
presence of a pyridine
such as an alkyipyridine...”



Regular Polysemy

- Ambiguity is not just for pyridine, but widespread throughout chemical nomenclature
- Some chemical terms are less ambiguous
 - e.g. “alkane”
 - No specific-compound sense
 - Usually in class-of-compounds sense
 - Also has part-of-compound sense
- Other regular polysemies exist, e.g.:
 - Metonymy
 - Gene/protein ambiguity

Guidelines

- Apply to pre-existing NE annotation
- Classification problem
 - Assign exactly one “subtype” to each NE
- Use informal “practise” rounds on other papers to develop guidelines
- Test agreement on 42 papers

Example

In addition, we have found in previous studies that the Zn²⁺-Tris system is also capable of efficiently hydrolyzing other β-lactams, such as clavulanic acid, which is a typical mechanism-based inhibitor of active-site serine β-lactamases (clavulanic acid is also a fairly good substrate of the zinc-β-lactamase from *B. fragilis*).

Example

In addition, we have found in previous studies that the Zn²⁺-Tris system is also capable of efficiently hydrolyzing other β-lactams, such as clavulanic acid, which is a typical mechanism-based inhibitor of active-site serine β-lactamases (clavulanic acid is also a fairly good substrate of the zinc-β-lactamase from *B. fragilis*).

EXACT

CLASS

PART

Subtypes for CM

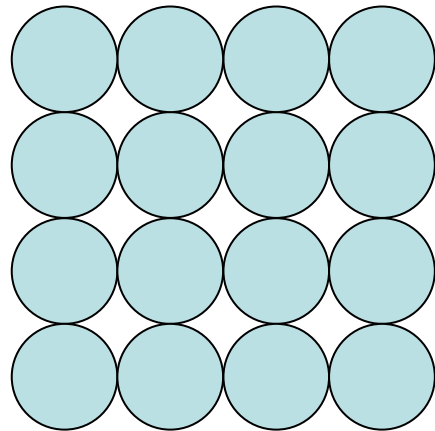
- EXACT Specific chemicals
- CLASS Classes of chemicals
- PART Parts of chemicals
- SPECIES “Atmospheric Carbon”
- SURFACE Surfaces
- POLYMER Polymers
- OTHER Very Rare

SPECIES

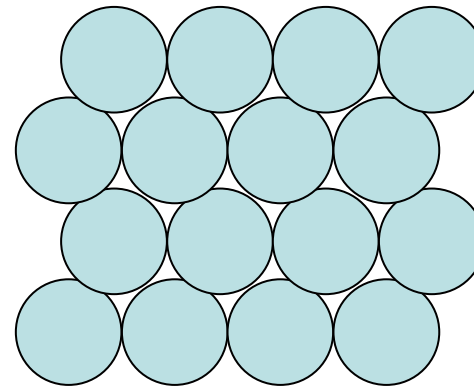
- “Atmospheric carbon”
 - Mostly in CO₂, not as soot
 - Carbon atoms as part of bulk matter, not part of individual molecular structures
 - 1kg atmospheric carbon = 3.67kg CO₂
 - Usage is more typical of EXACT than PART
- Elements ONLY
- Contexts for SPECIES:
 - Elemental analysis, ICP, XRF
 - Toxic elements (e.g. arsenic)
 - Environmental and metabolic cycles
- Conservation of number of atoms is often important

SURFACE

- Part of bulk matter, not a chemical structure
- Surface notations

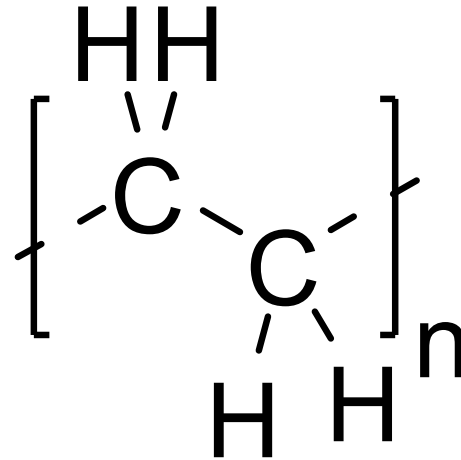


Ag(100)



Ag(111)

POLYMER



- Different samples of this polymer can have:
 - Different values, distributions of n
 - Different end groups
 - Different patterns of branching
- Yet *all* be called “polyethylene”

Compounds

- Compound nouns often contain a subtype-indicating head noun
 - “pyridine ring”
 - “methyl group”
 - “methyl compounds”
- In theory – hard to assign
 - “the ring as found in pyridine”
 - “the ring that defines the pyridines”
 - Redundant, like “tuna fish”, “pine tree”

Compounds

- Compound nouns often contain a subtype-indicating head noun
 - “pyridine ring”
 - “methyl group”
 - “methyl compounds”
- In theory – hard to assign
- For annotation – (usually) follow head noun

Inter-Annotator Agreement

- 42 papers, already annotated for NEs
- 2 annotators
 - Both PhD chemists
 - Both guidelines developers
- Reference to guidelines, reference sources etc.
- No conferring, or reference to previous attempts
- 86.0% Agreement
- Cohen's kappa = 0.784

Results By Subtype

Subtype	N (1 st annotator)	%	N (2 nd annotator)	%	F (%)
EXACT	3402	49.5	3246	47.3	89.9
CLASS	1114	16.2	1125	16.4	81.7
PART	1982	28.9	2118	30.9	84.3
SPECIES	233	3.4	194	2.8	77.3
SURFACE	73	1.1	131	1.9	63.7
POLYMER	58	0.8	49	0.7	74.8
OTHER	3	0.04	2	0.03	0.0

Automated Classification

- Motivation:
 - Investigate tractability
 - Establish “baseline” metrics
 - Keep it simple
- Straightforward classification task
 - Maximum Entropy classifier
- Absolute baseline – always EXACT
- Simple features

Feature Set

- The name itself
- Previous token
- Next token
- Suffix (4 characters)
- Plural (Ends in “s”)

Results

Features	Accuracy (%)		K	
None	49.5		0.0	
Name	56.2	+6.7	0.213	+0.213
Suffix	59.2	+9.7	0.303	+0.303
Plural	53.4	+13.9	0.114	+0.114
Previous token	54.2	+14.7	0.208	+0.208
Next token	61.0	+20.5	0.311	+0.311
All but name	67.3	-0.1	0.468	-0.002
All but suffix	67.0	-0.4	0.459	-0.011
All but plural	66.1	-1.3	0.447	-0.023
All but previous token	66.7	-0.7	0.452	-0.018
All but next token	62.0	-5.4	0.372	-0.098
All	67.4		0.470	

Conclusions

- We can reliably hand-annotate EXACT/CLASS/PART distinctions
- Automated annotation is tractable but with considerable room for improvement
- The next steps
 - Investigate deployment in IR systems
 - Investigate deployment in IE systems

Acknowledgements

- Peter Murray-Rust
- Royal Society of Chemistry
- UK eScience Programme
- EPSRC