

---

# Dictionary or Machine Learning based Chemical Named Entity Recognition?

---



**Fraunhofer** Institute  
Algorithms and  
Scientific Computing

Corinna Kolářik, Roman Klinger

6th Text Mining Symposium 2008 – Bonn, Germany

30 September 2008

---

# Outline

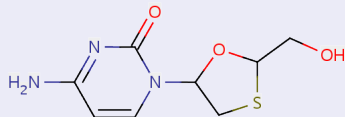
- 1 Introduction
- 2 Dictionary-based Named Entity Recognition
  - Resources
  - Test Corpus
  - Evaluation
  - Conclusion
- 3 Recognition of IUPAC and IUPAC-like Chemical Names
  - Machine Learning-based Named Entity Recognition
  - Corpora, Tokenization, and Features
  - Evaluation
  - Conclusion
- 4 Conclusion and Summary

# Introduction – Chemical Named Entity Recognition

... MT-4/DOX 500 cells showed reduced intracellular accumulation and retention of DOX and increased ATP-dependent rhodamine 123 efflux. The cells were also resistant to several anticancer agents such as mitoxantrone, 7-ethyl-10-[4-(1-piperidino)-1-piperidino]carbonyl oxycamptothecin, and 7-ethyl-10-hydroxycamptothecin. AZT was 7.5-fold less inhibitory to HIV-1 replication in MT-4/DOX 500 cells than in MT-4 cells. Furthermore, the anti-HIV-1 activity of lamivudine was severely impaired in MT-4/DOX 500 cells. ...

PMID 12488537: Wang X, et.al *Mol Pharmacol.* 2003

## Normalization: Lamivudine



- Pharm. Classes: Anti-HIV Agents, Reverse Transcriptase Inhibitors
- Drug Target: P protein, Gag-Pol polyprotein
- ...

# Dictionary-based Named Entity Recognition of Chemical Names

# Dictionary Example

## Normalization

DB00709

SMILES: CC(=O)NC(CS)C(=O)O

InChI=1/C8H11N3O3S/c9-5-1-2-11(8(13)  
10-5)6-4-15-7(3-12)14-6/h1-2,6-7,12H,  
3-4H2,(H2,9,10,13)/t6-,7+/m0/s1/f/h9H2

DB05246

SMILES: CC1(CC(=O)N(C1=O)C)C2=CC=CC=C2

InChI=1/C12H13NO2/c1-12(9-6-4-3-5-7-9)  
8-10(14)13(2)11(12)15-/h3-7H,8H2,1-2H3

## Synonyms

3TC; Epivir; Mucomyst; Epivir-HBV;  
Epzicom; Hepitec; Heptovir; Zeffix; 4-  
amino-1-[(2R,5S)-2-(hydroxymethyl)-  
1,3-oxathiolan-5-yl]pyrimidin-2-one;

...

Methsuximide; Petinutin; Celontin;  
1,3-dimethyl-3-phenylpyrrolidine-2,5-  
dione; ...

# Terminological Resources

## ■ Commercial Databases

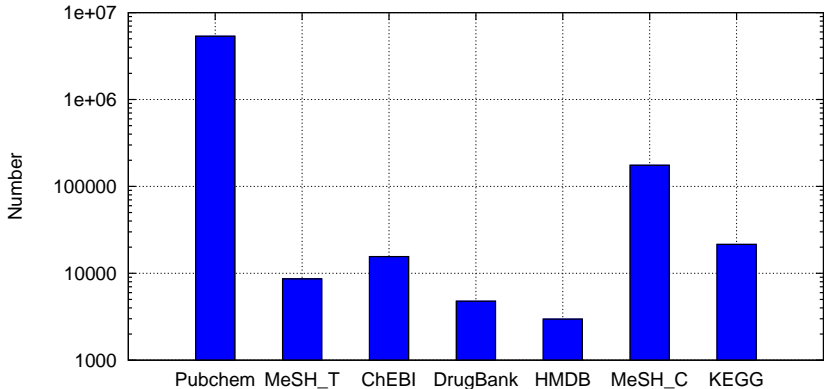
- CrossFire Beilstein, CAS Registry<sup>SM</sup>, World Drug Index

## ■ Freely Available Resources

- PubChem
- Kyoto Encyclopedia of Genes and Genomes (KEGG)
- DrugBank
- Human Metabolome Database (HMDB)
- Chemical Entities of Biological Interest (ChEBI)
- MeSH Medical Subject Headings (referred to as MeSH\_T)
- Supplementary Concept Records provided by National Library of Medicine (referred to as MeSH\_C)

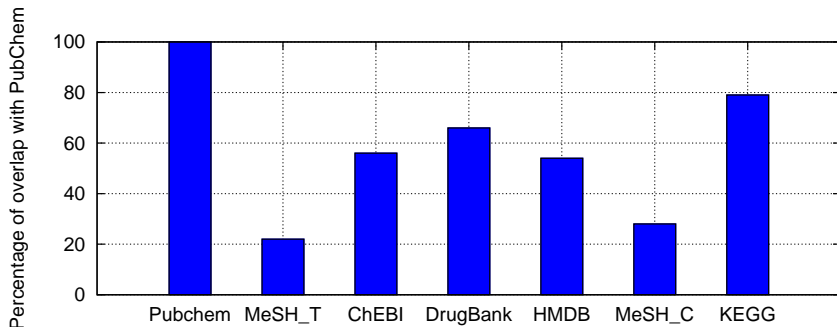
# Terminological Resources – Analysis

## Number of entries in extracted dictionaries



# Terminological Resources – Analysis

Are all synonyms included in PubChem?



Combining all analyzed dictionaries, **69 %** of the synonyms are not from PubChem but from the other resources.

# Evaluation of Resources

## How to analyze the usability of the resources for text mining?

- ⇒ **Test Corpus:** 100 Medline abstracts
- ⇒ **Assumption:** Some chemical term classes are to be found more easily in text than others.

# Annotation of a Test Corpus – Classes

## ■ TRIVIAL

- Single word terms (also if they were in fact IUPAC): **aspirin**, **estragon**, **testosterone**, **Acetylsalicylate**

## ■ IUPAC(-like)

- Multi-word systematic names:  
**N-substituted-pyridino[2,3-f]indole-4,9-dione**,  
**1-hexoxy-4-methyl-hexane**, **elaidic acid**,  
**1,4-dihydronaphthoquinones**

## ■ PART

- Partial chemical names (e. g. in enumerations):  
**8-(methylthio)-** and ..., **17beta-**

# Annotation of a Test Corpus – Classes

## ■ ABBREVIATION

- Abbreviations of names: TPA, AMPA
- Abbreviations as part of IUPAC names **not** tagged separately

## ■ SUM

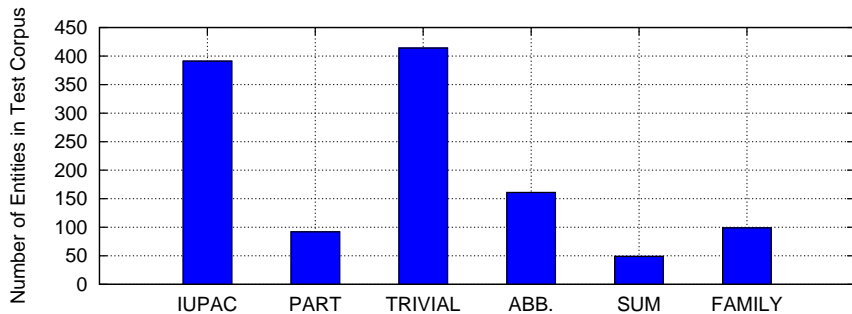
- Sum formulas: CH<sub>3</sub>SNa, KOH

## ■ FAMILY

- Chemical Families: disaccharide, pyrimidine, hydrazides
- **Not** pharmacological/functional families:  
(e.g.: anti-inflammatory drug, chelator)

# Annotation of a Test Corpus

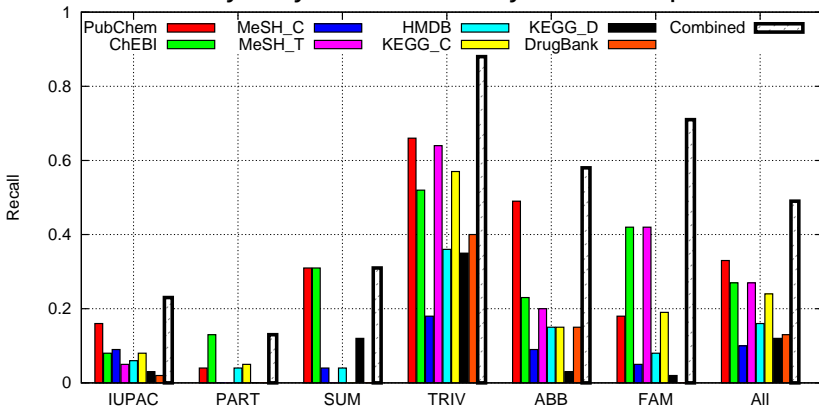
## Number of classes in test corpus



⇒ Altogether: **1222** entities

# Results on Test Corpus

Exact match of synonyms in dictionary to test corpus



# Resource Merging

Need for merging objects from different resources in an intelligent way

Strategy:

- Structural Information: **InChI**

**Lamivudine**: `InChI=1/C8H11N3O3S/c9-5-1-2-11(8(13)10-5)6-4-15-7(3-12)14-6/h1-2,6-7,12H,3-4H2,(H2,9,10,13)/t6-,7+/m0/s1/f/h9H2`

- Widely used references: **CAS numbers**
- Other database references
- If no identifier is available: Comparison of synonyms by approximate string match

# Problems with InChI

Different information content levels present in InChI from different resources

InChIs with 'Mobile H Perception' unchecked

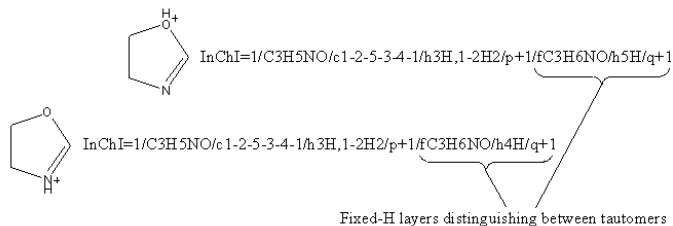
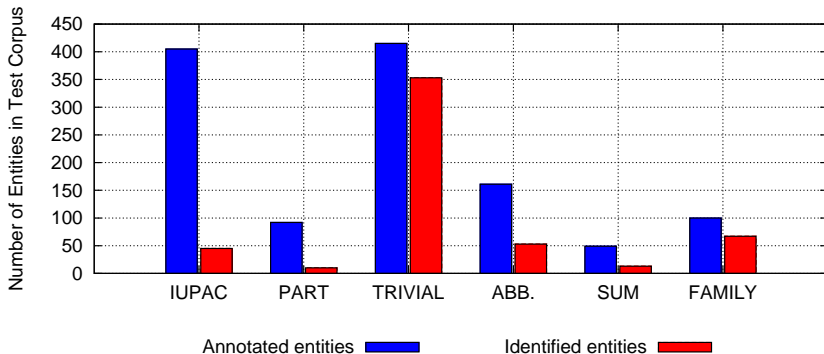


Figure taken from: <http://wwmm.ch.cam.ac.uk/inchifaq/>

# Results on Test Corpus

## Results after merging and curations with ProMiner



# Summary and Future Work

## Results of the Dictionary approach:

- Reasonable Recall for:
  - Trivial names,
  - Family names
- Things to do in the future:
  - Improve recognition by further curation efforts
  - Inclusion of further resources
- Recognition of IUPAC and IUPAC-like names is a problem

# Machine Learning-based Named Entity Recognition of IUPAC and IUPAC-like Names



# Introduction – Worst Abstract

“The worst abstract ever written, from a tokenization and coordination perspective”

Kevin B. Cohen at BioNLP mailing list (09/08/2008)

An efficient scalable synthesis of 2,3-epoxypropyl phenylhydrazones.

A series of mono and di-N-2,3-epoxypropyl N-phenylhydrazones have been prepared on a large scale by reaction of the corresponding N-phenylhydrazones of 9-ethyl-3-carbazolecarbaldehyde, 9-ethyl-3,6-carbazoledicarbaldehyde, 4-dimethyl-amino-, 4-diethylamino-, 4-benzylethylamino-, 4-(diphenylamino)-, 4-(4,4'-dimethyl-diphenylamino)-, 4-(4-formyldiphenylamino)- and 4-(4-formyl-4'-methyl-diphenyl-amino)benzaldehyde with epichlorohydrin in the presence of KOH and anhydrous Na<sub>2</sub>SO<sub>4</sub>.

PMID 17962747: Getautis V, et.al *Molecules*. 2006 Jan; 11(1):64–71.

chemical name

detected

missed by IUPAC tagger

found with dictionary

# Machine Learning for Text Mining

## Goal

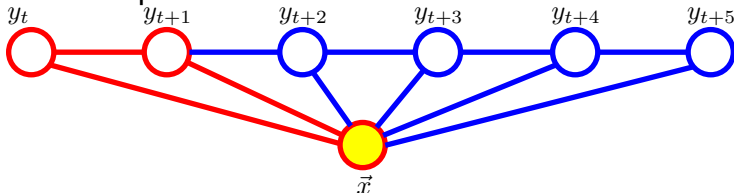
Map text sequence to label sequence:

$x =$	the	1	-	methylpropyl	is
$y =$	0	B-IUPAC	I-IUPAC	I-IUPAC	0

- Learning: ⇐ How to learn that model?  
Given training set  $E_t = \{(\vec{x}_1, \vec{y}_1), \dots, (\vec{x}_n, \vec{y}_n)\}$
- Question: ⇐ How get that from the model?  
Given new examples  $E = \{(\vec{x}_1, ?), \dots, (\vec{x}_m, ?)\}$

# Linear-Chain Conditional Random Field

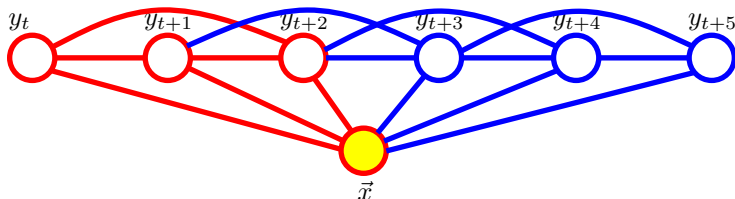
Probabilistic Graphical Model:



$$p_{\vec{\lambda}}(\vec{y}|\vec{x}) = \frac{1}{Z_{\vec{\lambda}}(\vec{x})} \cdot \prod_{j=1}^n \exp \left( \sum_{i=1}^m \lambda_i f_i(y_{j-1}, y_j, \vec{x}, j) \right)$$

$$Z_{\vec{\lambda}}(\vec{x}) = \sum_{\vec{y} \in \mathcal{Y}} \prod_{j=1}^n \exp \left( \sum_{i=1}^m \lambda_i f_i(y_{j-1}, y_j, \vec{x}, j) \right)$$

# 2nd-order Linear-Chain Conditional Random Field



$$p_{\vec{\lambda}}(\vec{y}|\vec{x}) = \frac{1}{Z_{\vec{\lambda}}(\vec{x})} \cdot \prod_{j=1}^n \exp \left( \sum_{i=1}^m \lambda_i f_i(y_{j-2}, y_{j-1}, y_j, \vec{x}, j) \right)$$

$$Z_{\vec{\lambda}}(\vec{x}) = \sum_{\vec{y} \in \mathcal{Y}} \prod_{j=1}^n \exp \left( \sum_{i=1}^m \lambda_i f_i(y_{j-2}, y_{j-1}, y_j, \vec{x}, j) \right)$$

# Corpus Generation

- **Training Corpus:** 463 **Medline** abstracts  
(selected with intermediate system and hand selected)
  - 161591 tokens, 3700 sentences, 3712 IUPAC entities
- **Test Corpus:** 1000 **Medline** abstracts, sampled
  - 124122 tokens, 5305 sentences, 151 IUPAC entities
- **Test Corpus:** Hand-Selected Sentences from 26 **patents**
  - 4309 tokens, 152 sentences, 411 IUPAC entities

# Features

## How to represent the text tokens?

- All Caps AMPA
- Real Number 0.142
- Is Dash - - - -
- Is Bracket {}[]
- Is Quote ' " "
- Is Slash / \
- Prefixes/Suffixes of different length pyrimidine
- Bag of Words pyrimidine
- Preceding/Succeeding White Space \_pyrimidine\_

# Features

- 3 Lists extracted from PubChem

- 1 714 prefixes of intermediate tokens of IUPAC Names
- 2 6161 suffixes of intermediate tokens of IUPAC Names
- 3 300 suffixes of last tokens of IUPAC Names

⇒ Generalize in exceed of training data

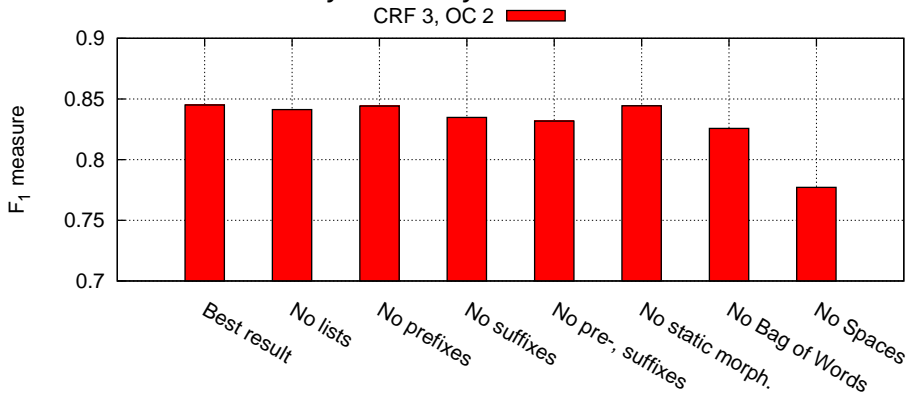
- Contextual information: Offset conjunction

(2R,10S)-N(1)-cyclopropylmethyl-2,10-dihydroxy-...

⇒ Leading to up to 521179 features!

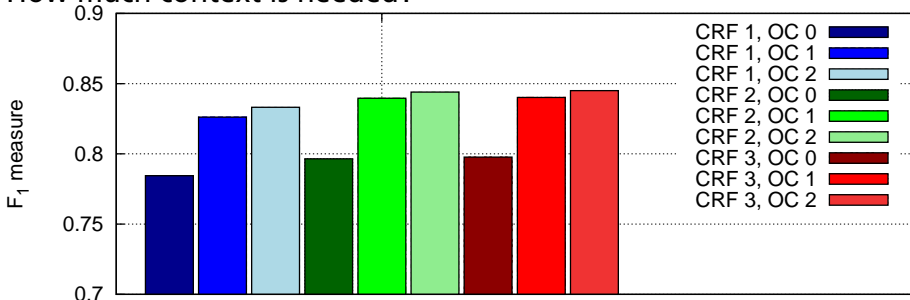
# Model Selection via Bootstrapping

Which Features are really necessary?

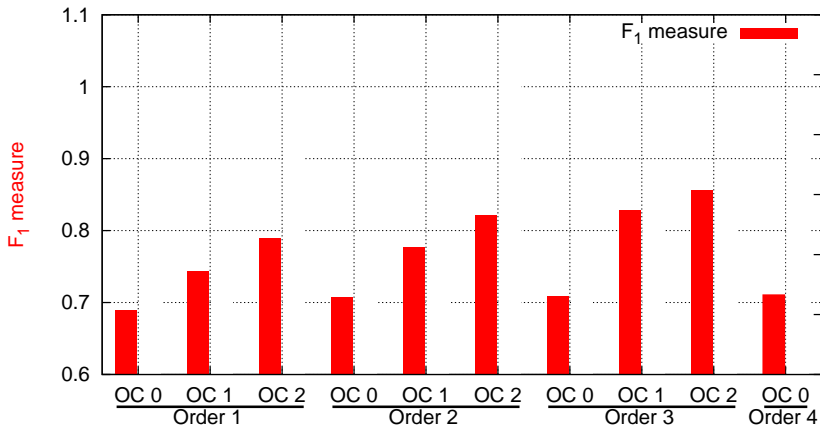


# Model Selection via Bootstrapping

How much context is needed?



# Results on Test corpus sampled from Medline



# Results: Summary

- **Medline Test Corpus: 85.6 %** (86.5 % P., 84.8 % R.)
- **Patent Test Corpus: 81.5 %** (77.2 % P., 86.4 % R.)
- **Test Corpus with several entities:**
  - On IUPAC and PART:  
 $F_1$ : **77.06 %** (**81.38 %** Precision, **73.18 %** Recall)
  - On all entities:  
**91.41 %** Precision, **29.04 %** Recall

# Conclusion

- Machine-Learning based system reasonable for IUPAC names
- High Performance possible on different Corpora
- Problems:
  - Trade-off between Durations and  $F_1$  measure
  - Problem: e.g. Enzymes: 2-phospho-D-glyceratehydro-lyase
- Improve Performance in  $F_1$  measure
  - Post-Processing
  - Extension of Training Corpus with Active Learning
- Improve Speed and Understandability of trained models

# Conclusion and Summary

- Rich resources for Dictionary Generation have been merged and applied for NER
- Good recognition of most classes
- Normalization straight-forward
  
- High Performance reached for IUPAC-like names with Machine Learning
- Detection of not fully enumerable names

⇒ Combination of different approaches

Thank YOU for your attention!

- Roman Klinger, Corinna Kolarik, Juliane Fluck, Martin Hofmann-Apitius, and Christoph M. Friedrich. **Detection of IUPAC and IUPAC-like Chemical Names.** *Bioinformatics*, 24(13):i268-i276, 2008. Proceedings of the International Conference Intelligent Systems for Molecular Biology (ISMB).
- Corinna Kolarik, Roman Klinger, Christoph M. Friedrich, Martin Hofmann-Apitius, and Juliane Fluck. **Chemical Names: Terminological Resources and Corpora Annotation.** In *Workshop on Building and evaluating resources for biomedical text mining* (6th edition of the Language Resources and Evaluation Conference), Marrakech, Morocco, 2008.

Corpus Download:

<http://www.scai.fraunhofer.de/chem-corpora.html>

[roman.klinger@scai.fhg.de](mailto:roman.klinger@scai.fhg.de)

<http://www.scai.fhg.de/klinger.html>

[corinna.kolarik@scai.fhg.de](mailto:corinna.kolarik@scai.fhg.de)

<http://www.scai.fhg.de/kolarik.html>