Chemical Names: Terminological Resources and Corpora Annotation

Corinna Kolářik[*†], Roman Klinger[†], Christoph M. Friedrich[†], Martin Hofmann-Apitius[*†], and Juliane Fluck[†]

[†]Fraunhofer Institute Algorithms and Scientific Computing (SCAI) Department of Bioinformatics Schloß Birlinghoven 53574 Sankt Augustin, Germany

[*]Bonn-Aachen International Center for Information Technology (B-IT) Department of Applied Life Science Informatics Dahlmannstrasse 2 D-53113 Bonn, Germany

corinna.kolarik@scai.fhg.de, roman.klinger@scai.fhg.de, christoph.friedrich@scai.fhg.de, martin.hofmann-apitius@scai.fhg.de, juliane.fluck@scai.fhg.de

Abstract

Chemical compounds like small signal molecules or other biological active chemical substances are an important entity class in life science publications and patents. The recognition of these named entities relies on appropriate dictionary resources as well as on training and evaluation corpora. In this work we give an overview of publicly available chemical information resources with respect to chemical terminology. The coverage, amount of synonyms, and especially the inclusion of SMILES or InChI are considered. Normalization of different chemical names to a unique structure is only possible with these structure representations. In addition, the generation and annotation of training and testing corpora is presented. We describe a small corpus for the evaluation of dictionaries containing chemical enities as well as a training and test corpus for the recognition of IUPAC and IUPAC-like names, which cannot be fully enumerated in dictionaries. Corpora can be found on `http://www.scai.fraunhofer.de/chem-corpora.html`

1. Introduction

In life science and chemical research a huge amount of new publications, research reports and patents is produced every year. High efforts were made to improve named entity recognition (NER) to support researchers to cope with the growing amount of publications. Analysis of the quality of developed methods have been focused to a great extend on the recognition of gene and protein names. Corpora for the main model organisms have been annotated and different systems have been evaluated in international assessments.

The identification of protein and gene names is still a challenge but as a result of the mentioned efforts, dictionary and rule based methods as well as machine learning techniques are now well established for protein and gene mentions in text. The Proceedings of the BioCreative II challenge (Hirschmann et al., 2007) give a good overview about the state-of-the-art methods and their performance.

A further important entity class is composed of small chemical compounds, for instance artificial substances, like drugs, or the organism's own biomolecules like metabolites or small signaling molecules. They are analyzed in many biological, medical or pharmacological studies to clarify their effect onto biological systems or to study the biological systems on its own.

In contrast to genes coded through a nucleotide sequence and protein macromolecules coded through amino acid sequences these small chemical molecules are represented in structures. InChI and SMILES are chemical structure descriptions that have been developed to refer to a compound with a unique textual compound identifier. In addition the largest commercial chemical database (CAS) provide for its whole chemical compound content unique CAS registry numbers (e.g. 50-78-2 for Aspirin). These numbers are are often used for normalization in the chemical community but they are proprietary and contain no structural information. Because of a limited readability of such specifications for humans, trivial names or drug trade names and the nomenclature published by the *International Union of Pure and Applied Chemistry* (IUPAC, (McNaught and Wilkinson, 1997)) is commonly applied (Eller, 2006) in text. Also combinations of the different types of names as well as abbreviations, especially of often used substances, are in use.

A number of systems deal with the entity class of chemical names, spanning from manually developed sets of rules (Narayanaswamy et al., 2003; Kemp and Lynch, 1998), grammar or dictionary-based approaches (Anstein et al., 2006; Kolářik et al., 2007; Rebholz-Schuhmann et al., 2007) to machine learning based systems (Sun et al., 2007; Corbett et al., 2007).

Semantic search, classification of recognized names, or structure and substructure searches are improved by normalizing the names to the corresponding structure. Chemical dictionaries containing structural representation allows for direct mapping of recognized names to the corresponding structure at the same time. Therefore one main task during the development of dictionary based systems is the generation of comprehensive resources providing synonyms and unique identifiers for the normalization of the entities of interest.

For other representations of chemical structures like SMILES, InChI or IUPAC names such an enumeration is only possible for the most common substances. The full chemical space cannot be enumerated. Therefore dictionary independent systems are necessary for the recognition of these names. For machine learning based systems as well as for system evaluation, the annotation of text corpora is another main challenge.

To our knowledge, no general overview or evaluation on publicly available terminology resources, like databases, covering chemical entities is available. In this work, we give a sur-

vey of different data sources, and evaluate the general usability of the contained chemical terminology for Named Entity Recognition. Unfortunately, none of the corpora used for the existing approaches mentioned above is publicly available for the evaluation and development of new methods. Therefore, we annotated new corpora and provide them publicly together with the annotation guidelines on http://www.scai.fraunhofer.de/chem-corpora.html.
IUPAC and IUPAC-like names have been identified with a machine learning approach that is based on Conditional Random Fields (Lafferty et al., 2001). Beside trivial names, these are used most often in publications and cannot be enumerated fully in dictionaries (more details can be found in (Klinger et al., 2008)). We discuss our experiences in the generation and annotation of the corpora and give a short overview on the results.

2. Terminological Resources

Entity recognition approaches that are based on dictionaries rely on comprehensive terminology resources containing frequently used synonyms and spelling variants. An example excerpt of an extracted dictionary is given in Table 1. As for proteins and genes, databases could be a valuable resource to obtain chemical named entities and their synonyms. In this section we give an overview on available data sources.
Until recently, when the academic community started to build information sources for biologically relevant chemical compounds, chemical information was only available from commercial databases. The most important and largest resources not freely available are the CAS REGISTRY[1], the CrossFire Beilstein[2] database, and the World Drug Index[3].
For a deeper analysis we focus on freely available resources basically used in biomedicinal research. These are databases with public chemical content, thesauri and an ontology that have been growing over the last years. We concentrate on entities belonging to the class of small organic molecules and drugs from the context of human studies. Some of them contain very specific information and others cover a broad chemical space. The database PubChem[4] (Wheeler et al., 2008), the ChEBI ontology[5] (Degtyarenko et al., 2008), and MeSH[6] represent sources for a broad chemical space.
The more specialized data sources DrugBank[7](Wishart et al., 2008) and KEGG Drug[8] (Kanehisa et al., 2008) were considered as drug terminology resources. KEGG Compound[9] and the Human Metabolome Database (HMDB)[10] (Wishart et al., 2007) have been chosen as terminology resources for metabolic substances.

This survey does not claim to give a complete overview of all available chemical information resources. There is a number of other databases and resources covering specialized chemical information and a broader chemical space, e.g. UMLS[11] (Nelson et al., 2002) implying MeSH, MedlinePlus[12], and ChemIDplus[13] (Tomasulo, 2002).

2.1. Commercial Databases

CrossFire Beilstein database is a large repository for information of over 10 million organic compounds, determining their bioactivity and physical properties, ascertaining the environmental fates and their reactions. Beside structural information the entities are associated with chemical and physical facts, bioactivity data, and literature references.

CAS REGISTRY^{SM} provided by CAS, is one of the largest databases of chemical substance providing information about more than 33 million organic and inorganic substances as well as over 59 million sequences. To each substance, a unique ID (CAS Registry Number) is assigned, generated by CAS to link between the various nomenclature terms as a kind of normalization. These IDs have long been used as reference to chemicals in other databases as well as in text.

The World Drug Index contains chemical and biomedical data for over 80,000 marketed and development drugs with internationally recognized drug names, synonyms, trade names, and trivial names. Each record has a chemical structure and is classified by drug activity, mechanism of action, treatment, manufacturer, synonyms, and medical information.

2.2. Freely available Resources

From all resources introduced in this section individual dictionaries have been created and evaluated on the EVAL corpus (see Section 5.1).

PubChem consists of three linked databases – *PubChem Substance*, *PubChem Compound*, and *PubChem BioAssay*. They are part of the NCBI's Entrez information retrieval system[14]. *PubChem Compound* contains 18.4 million entries of pure and characterized chemical compounds, structure information, SMILES, InChI, and IUPAC but no further synonyms. *PubChem Substance* provides 36.8 million entries with information about mixtures, extracts, complexes, and uncharacterized substances or proteins. It comprises synonyms in the form of trivial names, brand names, IUPAC, but no SMILES, and only few mappings to InChI names. For the chemical dictionary names and synonyms as well as the chemical structure information are needed. Therefore a PubChem subset dictionary was generated with all *PubChem Substance* entries containing names, synonyms and links to corresponding entries of *PubChem Compound* (5,339,322 records).

Chemical Entities of Biological Interest (ChEBI) is a freely available controlled vocabulary of small molecular

[1] http://www.cas.org/expertise/cascontent/registry/index.html
[2] http://www.beilstein.com/
[3] http://scientific.thomson.com/products/wdi/
[4] http://pubchem.ncbi.nlm.nih.gov/
[5] http://www.ebi.ac.uk/chebi/init.do
[6] http://www.nlm.nih.gov/mesh/meshhome.html
[7] http://drugbank.ca/
[8] http://www.genome.jp/kegg/drug
[9] http://www.genome.jp/kegg/compound
[10] http://hmdb.ca/
[11] http://www.nlm.nih.gov/research/umls/
[12] http://medlineplus.gov/
[13] http://chem.sis.nlm.nih.gov/chemidplus/
[14] http://www.ncbi.nlm.nih.gov/

i	id_i	S_i
1	DB06151	CC(=O)NC(CS)C(=O)O; InChI=1/C5H9NO3S/c1-3(7)6-4(2-10)5(8)9/h4,10H,2H2,1H3,(H,6,7)(H,8,9)/t4-/m0/s1/f/h6,8H; Acetylcysteine; ACC; Mucomyst; Acetadote; Fluimucil; Parvolex; Lysox; Mucolysin; (2R)-2-acetamido-3-sulfanylpropanoic acid; ...
2	DB05246	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; Methsuximide; Petinutin; Celontin; 1,3-dimethyl-3-phenylpyrrolidine-2,5-dione; ...

Table 1: Example for a dictionary based on DrugBank, usually incorporated in rule based Named Entity Recognition systems. The identifier (in this case a DrugBank identifier) is denoted with id_i, the set of synonyms with S_i.

entities that intervene in the processes of living. Entities are organized in an ontological classification and are grouped by their chemical structure and functional properties. General chemical class terms, biological and pharmacological functions, and compounds with general names are covered as well as synonyms of the form of trivial name, IUPAC, and sum formula. For most of the chemical compounds SMILES and InChI names are given. We used the release version 35 of ChEBI provided in the OBO-format.

Medical Subject Headings (MeSH) is a controlled vocabulary thesaurus from the National Library of Medicine (NLM)[15]. It is used by NLM for indexing articles from the MEDLINE PubMED database as well as a catalog database for other media of the library. The terms are organized in a hierarchy to which synonyms as well as inflectional term variants are assigned. A subset of the MeSH thesaurus (version 2007 MeSH) covering the chemical category of MeSH (tree concepts with node identifiers starting with 'D') was extracted to give one dictionary of MeSH (referenced further as MeSH_T). Furthermore, NLM provides a compound list with over 175,000 entries containing synonyms like trivial and brand names, IUPAC and abbreviations which was used to generate another dictionary, referenced further as MeSH_C.

Kyoto Encyclopedia of Genes and Genomes (KEGG) is a composite database that integrates genomic, chemical, and systemic functional information. Two sub-databases – *KEGG COMPOUND* and *KEGG DRUG* – are considered to be terminology resources for the dictionary creation. The types of compounds provided by *KEGG COMPOUND* span from single ions (e.g. Mg^{2+}), simple compounds (like different sugars or cofactors of enzymes, metabolites, products of microorganisms, or nuclear receptor compounds like GW 6471) to peptides and basic RNAs – all essential endogenous molecules of cells. *KEGG DRUG* covers all approved drugs in the United States of America and Japan. Every entry of both databases is linked to a unique chemical structure and to standard generic names that could be of the type IUPAC and trivial name. For the creation of the two dictionaries KEGG_C and KEGG_D the fields 'NAME', 'FORMULA', and 'DBLINKS' of the KEGG proprietary format files *compound* and *drug* have been used.

DrugBank is a specific database about pharmaceuticals, that combines detailed chemical, pharmacological, and pharmaceutical information with drug target information. It provides trivial, brand, and brand mixture names, IUPAC and a structure for almost every entity as SMILES or InChI. DrugBank is available as a single file in a proprietary format. Following fields have been extracted: 'Name', 'Synonyms', 'Brand Names', 'Brand Mixtures', 'Chemical IUPAC Name', 'Chemical Formula', 'InChI Identifier', 'Isomeric SMILES', 'Canonical SMILES', and 'CAS Registry Number'.

[15] http://www.nlm.nih.gov/

Resource	Number of entries
CrossFire Beilstein	10 mio.
CAS	33 mio.
World Drug Index	80,000
PubChem_C; PubChem_S	18.4 mio.; 36.8 mio.
MeSH_T	8,612
MeSH_C	175,136
ChEBI	15,562
KEGG (K-C; K-D)	21,498 (15,033; 6,834)
DrugBank	4,764
HMDB	2,968

Table 2: Total number of entities contained in chemical information resources (PubChem_C: PubChem Compound; PubChem_S: PubChem Substance; K-C: KEGG-compound; KEGG-Drug)

Human Metabolome Database (HMDB) is a freely available database containing detailed information about small molecule metabolites found in the human body. The focus lies on quantitative, analytic or molecular scale information about metabolites, their associated enzymes or transporters and their disease-related properties. The database currently contains nearly 3000 metabolite entries, like hormones, disease-associated metabolites, essential nutrients, and signaling molecules as well as ubiquitous food additives and some common drugs. HMDB is downloadable as a single file with a similar proprietary format as DrugBank. Following fields have been extracted: 'Name', 'Common Name', 'Synonyms', 'Chemical IUPAC Name', 'Isomeric SMILES', 'Canonical SMILES', 'InChI Identifier', and 'CAS Registry Number'.

3. Analysis of the Chemical Information Resources

In this section we discuss the general usability of the above mentioned resources for dictionary based named entity recognition approaches. The resources were analyzed with

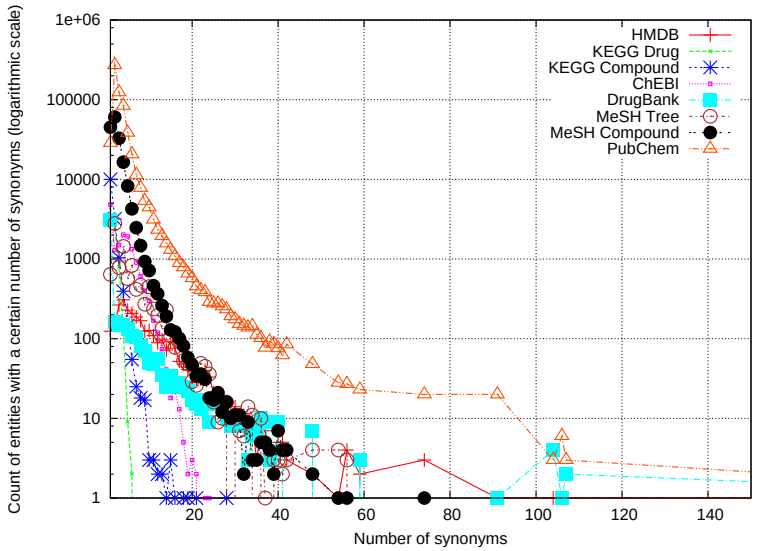

Figure 1: Plot of the synonym count distribution for the analyzed databases

	PubChem subset	MeSH_T	ChEBI	DrugBank	HMDB	MeSH_C	KEGG
SMILES	4,080,909	—	8,371	4,489	2,881	—	—
InChI	4,080,909	—	8,280	4,486	2,859	—	17,021
CAS	397,858	—	4,566	2,223	2,527	175,136	13,545
Percentage of synonyms covered by PubChem	100 %	22 %	56 %	66 %	54 %	28 %	79 %
Cross references	yes	—	yes	yes	yes	—	yes
Total No. of entries	5,339,322	8,612	15,562	4,764	2,968	175,136	21,498

Table 3: Overview of the linkage of the entities to structure information in the analyzed data sources. For PubChem only *PubChem Substance* entries containing a *PubChem Compound* link were included. For KEGG the respective values of the drug and compound sub-databases were unified. For the PubChem coverage all synonyms of all entries are compared.

regard to following properties:

- Total number of entries,
- Provided number of synonyms,
- Linkage to a structure, and
- Cross linkage to other databases.

Table 2 gives an overview about the total amount of the entities provided by the analyzed sources. All commercial databases contain a huge number of chemical entities, reflecting their growth for a long time. In comparison to them, PubChem is the biggest collection of public chemical data, followed by MeSH compounds. The remaining specialized resources, like DrugBank or HMDB, contain fewer entities but highly comprehensive biomedical information about them. Figure 1 reflects the distribution of the occurrences of synonyms for every analyzed resource. Most entries contain only few synonyms. Entries of PubChem, both MeSH dictionaries, and DrugBank as well as HMDB contain a high amount of synonyms. A high number of provided synonyms is of high value for the creation of the dictionaries. A comprehensive coverage of the chemical terms and their synonyms used in text leads to a good performance of a dictionary-based NER approach by avoiding a high false negative rate. Comparison of the synonyms contained in PubChem to the other databases (cf. Table 3) showed that there are differences in the synonym coverage in the analyzed resources. About 79 % of the KEGG synonyms are included in PubChem and 55 % of the CheBI entities but only 22 % of the MeSH tree synonyms could be found. Combining all analyzed dictionaries, 69 % of the synonyms are not from PubChem but from the other resources. Hence, it is meaningful to use an all-integrating dictionary. instead of incorporating only PubChem.

Table 3 presents the number of the resource which are mapped to InChI, SMILES or CAS. Unique representations are relevant for the mapping and normalization of the identified chemical names from text to a chemical structure. All entries in the selected PubChem subset contain InChI information and two third of the entries contain the CAS registry number. Most entries in DrugBank, HMDB, ChEBI are mapped to all three chemical representations and in KEGG, InChI and CAS registry numbers are included. All entries of MeSH_C are mapped to CAS identifiers but no other chemical representations like InChI is given.

In addition no cross references to other data sources are included. The other sources contain a high number of cross

references and references to PubChem, KEGG and ChEBI are given in all databases. PubChem contains the highest number of cross references and in addition links to MeSH.

4. Annotation and Corpus Generation

For evaluation purposes of NER-systems as well as for the training of machine learning based methods annotated corpora are needed. Corbett et al. (2007) describe a corpus annotation, but the corpus as well as the annotation guidelines are not publicly available. Because annotated corpora for the chemical domain are not public available yet, we describe three corpora consisting of MEDLINE abstracts. A small evaluation corpus (EVAL corpus) containing entities of all classes described in Tables 3 and 4.1 has been annotated to give an overview of the different chemical name classes found in MEDLINE text. This corpus will be used for a first assessments of chemical dictionaries and for the evaluation of methods for chemical name recognition. In addition, a training and a test corpus was generated for the machine learning based recognition of IUPAC and IUPAC-like names and has been annotated with the classes IUPAC and PART. In the following sections our assignment of chemical terms to various defined annotation classes and the corpus annotation is described.

4.1. Chemical Entity Classes used for the Annotation

To allow an annotation even for non-chemical experts a simplified classification schema with respect to chemical classification was developed. The defined classes are IUPAC, PART, TRIVIAL, ABB, SUM, and FAMILY, shown in Table 3 with descriptions and examples. The separation between TRIVIAL and IUPAC names is based on the term length, names with only one word were classified as TRIVIAL even if they were IUPAC names. Multi word systematic and semi-systematic names are always annotated as IUPAC. This includes names that imply only a IUPAC-like part (e.g. 17-alpha-E) or names including a labeling (e.g. 3H-testosterone). This does not follow strictly the definition of IUPAC, but such terms are less likely contained in databases and cannot be found with a pure dictionary-based approach. For the correct resolution of enumerations, partial chemical names have been annotated separately as PART, but chemical names were not tagged in other entities (e.g. in protein names). Names were only tagged as FAMILY if they describe well defined chemical families but not pharmacological families (e.g. glucocorticoid was labeled but not anti-inflammatory drug). Substances used as base for building various derivates and analogs were tagged as IUPAC, not as FAMILY (e.g. 1,4-dihydronaphthoquinones). More examples and their labels used for the annotation are provided for clarification in Table 4.1. All defined classes were used for the annotation of the evaluation corpus. This annotation allows the assessment of distribution of chemical names in MEDLINE text and the coverage of the different dictionaries and recognition approaches. We do not imply to use this classification as final annotation scheme for chemical name annotation. Further iterations of evaluation and annotation are necessary and are work in progress including more chemical experts.

4.2. Corpus Selection for the Annotation and Evaluation of all Chemical Classes

Based on the assumption that abstracts containing IUPAC names also contain other nomenclatures, a preliminary system for detecting IUPAC names as described in Section 4.3 (Klinger et al., 2008) was applied to select abstracts from MEDLINE containing at least one found entity. Next to abstracts selected with this procedure, we selected abstracts containing problematical cases as well as those containing no entities. This procedure formed a corpus of 100 abstracts containing 391 IUPAC, 92 PART, 414 TRIVIAL, 161 ABB, 49 SUM, and 99 FAMILY entities.

4.3. Corpus Generation for the Recognition of IUPAC and IUPAC-like Entities

As a training corpus for a Conditional Random Field (CRF), 463 abstracts have been selected from 10,000 sampled abstracts from MEDLINE. It was annotated by two independent annotators. A conclusive training corpus was generated using a combination of both annotations by an independent person. This resulted in a corpus containing 161,591 tokens with 3,712 IUPAC annotations. Here, the class PART was included in the class IUPAC due to morphological similarity of these classes which is important for the machine learning approach described in Section 5.2.
A test corpus was selected to test the system trained on the above described training corpus. For that, 1000 MEDLINE records with 124,122 tokens were sampled equally distributed from full MEDLINE and has been annotated. It comprises 151 IUPAC entities. The sampling process ensures to have representative text examples of the full MEDLINE. This is especially beneficial for a correct analysis of the false positives.

4.4. Inter-Annotator Agreement

For the corpus with all chemical entities described in Section 4.2 and the training corpus described in Section 4.3, the inter-annotator agreement was evaluated.
Recognizing the boundaries without considering the different classes on the test corpus described in Section 4.2, the inter-annotator F_1 is 80 % and for the IUPAC entity in the training corpus, the F_1 measure is 78 %. For both corpora conclusive corpora were generated. The conclusive training corpus and the first-annotated corpus differ to a lower degree, the inter-annotator F_1 measure is 94 %. In contrast Corbett et al. (2007) claimed 93 % for the training corpus for the system *OSCAR*. One reason for the lower F_1 measure in the first annotation in comparison to the result of Corbett and his colleagues is our differentiation of the IUPAC entity to other chemical mentions. The appropriate usage of those is not always easy to decide, while all chemical mentions in the corpus generated by Corbett are combined in one entity (see Section 5.2 for more details). Another reason is the different experience level of our annotators. One annotator participated in the development of the annotation rules. The corpus was annotated partly by this person more than once during this process. The second person annotated the whole set based on the annotation guideline without an intermediate revision. Therefore, we propose a two step process

Chemical Class	Description	Example Annotation
IUPAC	IUPAC names, IUPAC-like names, systematic, and semi-systematic names	1-hexoxy-4-methyl-hexane
PART	partial IUPAC class names	17beta-
TRIVIAL	trivial names	aspirin, estragon
ABB	abbreviations and acronyms	TPA
SUM	sum formula, atoms, and molecules, SMILES, InChI	KOH
FAMILY	chemical family names	disaccaride

Table 4: Chemical entity classes used for the corpora annotation

Name	Labeled Sequence	Label	Explanation
Acetylsalicylate	Acetylsalicylate	TRIVIAL	
elaidic acid	elaidic acid	IUPAC	multi word systematic and semi systematic names are labeled as IUPAC
testosterone	testosterone	TRIVIAL	
3H-testosterone	3H-testosterone	IUPAC	contains part IUPAC-like structure (3H-);
17-alpha-E	17-alpha-E	IUPAC	E = chemical abbreviation
17beta-HSD	—	—	HSD = protein name
N-substituted-pyridino[2,3-f] indole-4,9-dione	N-substituted-pyridino[2,3-f] indole-4,9-dion	IUPAC	
2-acetyloxybenzoic acid	2-acetyloxybenzoic acid	IUPAC	
Ethyl O-acetylsalicylate	Ethyl O-acetylsalicylate	IUPAC	
pyrimidine	pyrimidine	FAMILY	
1,4-dihydronaphthoquinones	1,4-dihydronaphthoquinones	IUPAC	
Ca(2+)	Ca(2+)	SUM	
(14)C	(14)C	SUM	

Table 5: Annotation examples

for further annotations. In a first step an inter-annotator agreement should be build only on a small set of annotated abstracts and discrepancies could be reviewed with all annotators. Then the larger set of abstracts could be annotated with higher confidence.

5. Recognition of Chemical Names

5.1. Dictionary-based Recognition of Chemical Compounds

Dictionaries built from the different terminological resources were used to recognize chemical entities in the EVAL corpus. Following constraints were used for all searches:

- No curation of the created raw dictionaries was done, which means that no names were removed, added or changed.
- All synonyms were searched with a simple case insensitive string search, dashes were ignored.
- No control of the correct association of the found names to the corresponding entry was performed.

The results with uncurated dictionaries and such a simple search strategy should only give a rough estimate of the coverage of different sources and the efforts which have to be invested in curation and search strategies.
The search results obtained with every individual dictionary and a combination of the results of all dictionaries are provided in Table 6. The first two rows show precision and recall on a combination of all annotation classes. The rates in brackets were obtained when also partial matches were considered as true positives. The highest precision rates were achieved by the KEGG Drug dictionary (59 %) followed by the MeSH_C dictionary (44 %). The lowest precision of 13 % and 15 % was obtained by ChEBI and PubChem respectively. Many unspecific terms are contained in ChEBI (e.g. groups or inhibitors), and also terms that have not been annotated as a chemical family term (e.g. enzyme inhibitors or adrenergic agonist). Such terms were considered to be pharmaceutical property terms. Additionally, many other names are unspecific, like one-character tokens (e.g. D, J) and common word names (e.g. at, all). Therefore, we conclude that curation processes are necessary to achieve a higher performance with the dictionaries. Experiences with the gene and protein name recognition (Hirschmann et al., 2007) let us assume that the precision could be highly enhanced through dictionary curation and more elaborate named entity recognition techniques.

The recall of the dictionary based named entity recognition is low. The highest recall was obtained with the PubChem dictionary identifying 33 % of all entries, followed by the ChEBI and the MeSH_T dictionary (both 27 %). The conflation of all search results enhances the recall to 49 %, but decreases the precision to 13 %. The participation of the different dictionaries on the combined result has to be checked further for recall and precision.

Class	PubChem	ChEBI	MeSH_C	MeSH_T	HMDB	KEGG_C	KEGG_D	DrugBank	Combined
ALL (1206)	*0.15 (0.26)* 0.33 (0.60)	*0.13 (0.34)* 0.27 (0.68)	*0.44 (0.64)* 0.10 (0.15)	*0.34 (0.42)* 0.27 (0.34)	*0.21 (0.44)* 0.16 (0.33)	*0.30 (0.54)* 0.24 (0.43)	*0.59 (0.76)* 0.12 (0.16)	*0.33 (0.43)* 0.13 (0.17)	*0.13 (0.22)* 0.49 (0.85)
IUPAC (391)	0.16 (0.69)	0.08 (0.85)	0.09 (0.21)	0.05 (0.29)	0.06 (0.44)	0.07 (0.51)	0.03 (0.17)	0.01 (0.17)	0.23 (0.94)
PART (92)	0.04 (0.32)	0.13 (0.72)	0.00 (0.05)	0.00 (0.01)	0.04 (0.32)	0.05 (0.24)	0.00 (0.00)	0.00 (0.00)	0.13 (0.75)
SUM (49)	0.31 (0.73)	0.31 (0.88)	0.04 (0.08)	0.00 (0.00)	0.00 (0.30)	0.12 (0.46)	0.00 (0.00)	0.00 (0.00)	0.31 (0.88)
TRIV (414)	0.66 (0.82)	0.52 (0.78)	0.18 (0.19)	0.64 (0.65)	0.36 (0.42)	0.57 (0.64)	0.35 (0.36)	0.40 (0.41)	0.88 (0.97)
ABB (161)	0.49 (0.72)	0.23 (0.55)	0.09 (0.11)	0.2 (0.23)	0.15 (0.34)	0.15 (0.32)	0.03 (0.03)	0.03 (0.03)	0.58 (0.83)
FAM (99)	0.18 (0.5)	0.42 (0.69)	0.05 (0.09)	0.42 (0.42)	0.08 (0.13)	0.19 (0.35)	0.17 (0.03)	0.00 (0.03)	0.71 (0.89)

Table 6: Comparison of the entities found in the evaluation corpus with dictionaries based on the analyzed resources. All annotation classes are considered. (The total number of the annotated entities per class are given in brackets.) Precision (slanted) and recall are given for an exact match of an entity and a match where the identification of a subset of the term is sufficient (values behind the recall values in brackets).

The analysis of the recall for every single annotation class confirms our hypothesis that names belonging to the TRIVIAL class could be found with the highest recall. The search with the PubChem dictionary identified 66 %, followed by MeSH_T with 64 % and KEGG_C with 57 %. The combination of the results lead to a promising recall of 88 %. Considering the recognition of family names by the ChEBI and the MeSH_T dictionary obtained the highest value (both 42 %). This is not very remarkable, because only those two resources contain general chemical group and family terms in their hierarchy. Sum formulas (mainly annotated as shown in Table 4.1) were only recognized to a certain degree by ChEBI, PubChem (both 31 %), and KEGG_C dictionary (12 %). The recognition rate of the ABB class has to be taken with caution because abbreviations are often short names, sometimes only one character long and therefore highly ambiguous.

As we previously assumed, IUPAC names have been recognized with a low recall by all tested dictionaries. The partial match rate is high, especially for the PubChem and ChEBI dictionary. Some partial matches, e.g. 'testosterone' in '3H-testosterone', could be accepted, but many terms, e.g. diethyl or benzoyl being part of 'diethyl N-[2-fluoro-4-(prop-2-ynylamino)benzoyl]-L-glutamate', increase the rate of false positive partial matches. Therefore, strategies need to be integrated for an efficient recognition system to avoid such problems.

In summary we can conclude from this experiment that the recall of a simple search strategy that uses the individual uncured dictionaries is low. The combination of all dictionaries leads to an acceptable rate for TRIVIAL and FAMILY names but not for IUPAC and PART names. For the recognition of the latter two a machine learning approach might be advantageous compared to a dictionary approach. Thus a machine learning based strategy for the IUPAC name recognition is described in the next section.

	Precision	Recall
IUPAC tagger on test corpus sampled from MEDLINE (IUPAC + PART entities)	86.50	84.80
IUPAC tagger on EVAL corpus (all entities)	91.41	29.04
IUPAC tagger on EVAL corpus (IUPAC + PART entities)	81.38	73.18
IUPAC tagger on EVAL corpus (IUPAC entities)	—	77.11
IUPAC tagger on EVAL corpus (PART entities)	—	41.18
IUPAC tagger on EVAL corpus (TRIVIAL entities)	—	8.42
OSCAR on EVAL corpus with all entities	52.09	71.86

Table 7: Results of the machine learning-based tagger and of the system *OSCAR* for IUPAC entities and all entities on the EVAL corpus and on the test corpus sampled from MEDLINE (in %).

5.2. CRF-based IUPAC Name Recognition

To improve the recognition of IUPAC names, the training corpus described in Section 4.3 was used to train a Conditional Random Field. Due to the morphological similarity of IUPAC and PART entities they have been combined leading to a system that does not separate these two classes. The parameter optimization (e.g. feature selection) is described in detail in Klinger et al. (2008).
An evaluation on the sampled test corpus of 1000 abstracts from MEDLINE shows an F_1 measure of 85.6 % with a pre-

cision of 86.5 % and a recall of 84.8 %. Applying this tagger on the EVAL corpus with several entity classes described in Section 4.2, it recognizes 73.18 % of the IUPAC and IUPAC-like names with a precision of 81.38 % (considering only IUPAC and PART names as true positive hits). The recall on the separated classes IUPAC and PART (namely 77.11 % and 41.18 %) on the EVAL corpus motivates the combination of these classes for machine learning purposes.

The precision of 91.41 % on all entities is much higher than only on the IUPAC entities due to the recognition of 8.42 % of the TRIVIAL class names. They are frequently used within IUPAC terms and cannot be easily separated by the system. A separation from the other classes ABB, SUM, and FAMILY is perfectly given.

It needs to be analyzed if trivial names could be recognized with a machine learning based method with similar performance to enhance the recall of the system which is now at 29 % considering all chemical classes. Here, an additional annotation of the training set is necessary.

To compare the *OSCAR* software, this approach was also used for the recognition of all entities in the EVAL corpus. *OSCAR* has an overall high recall of almost 72 % accompanied with a precision of 52 %. The recall is similar to the reports in (Corbett et al., 2007) (73.5 % recall, 75.3 % precision) but the precision is lower. We did not analyze the results in detail but certainly one reason for the lower precision can be found in the different annotation of chemical entities underlying the training corpus used in *OSCAR*. One difference is for example the annotation of more general annotation of chemical names (e.g. dry ice).

6. Conclusion

To a certain amount, trivial names and family names but not IUPAC like names are covered by the different chemical resources analyzed in this paper. PubChem, as the largest resource, does not include all names covered by the smaller sources. Hence, the combination of the search results from all terminologies lead to a high increase in recall, especially for family and trivial names. The development of a training corpus for IUPAC like entities lead to a performant CRF-based IUPAC tagger.

These results are motivating for further investigations in the generation of dictionaries as well as testing different annotation classes to be used for training and the combination of machine learning based chemical name recognition and dictionary based normalization of chemical names.

7. Acknowledgments

We would like to thank Theo Mevissen for dictionary generation and technical support and Harsha Gurulingappa and Erfan Younesi for corpus annotation. The work of Roman Klinger is funded by the MPG-FhG Machine Learning Collaboration (`http://lip.fml.tuebingen.mpg.de/`).

8. References

S. Anstein, G. Kremer, and U. Reyle. 2006. Identifying and classifying terms in the life sciences: The case of chemical terminology. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bene Maegaard, Joseph Mariani, Jan Odijk, and Dnaiel Tapias, editors, *Proc. of the Fifth Language Resources and Evaluation Conference*, pages 1095–1098, Genoa. Italy.

P. Corbett, C. Batchelor, and S. Teufel. 2007. Annotation of chemical named entities. In *BioNLP 2007: Biological, translational, and clinical language processing*, pages 57–64, Prague, June.

K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, and M. Ashburner. 2008. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, 36(Database issue):D344–D350, Jan.

G. A. Eller. 2006. Improving the quality of published chemical names with nomenclature software. *Molecules*, 11:915–928.

L. Hirschmann, M. Krallinger, and A. Valencia, editors. 2007. *Proc. of the Second BioCreative Challenge Evaluation Workshop*. Centro Nacional de Investigaciones Oncologicas, CNIO.

M. Kanehisa, M. Araki, S. Goto, M. Hattori, et al. 2008. Kegg for linking genomes to life and the environment. *Nucleic Acids Res*, 36(Database issue):D480–D484, Jan.

N. Kemp and M. Lynch. 1998. The extraction of information from the text of chemical patents. 1. identification of specific chemical names. *Journal of Chemical Information and Computer Sciences*, 38(4):544–551.

R. Klinger, C. Kolářik, J. Fluck, M. Hofmann-Apitius, and C. M. Friedrich. 2008. Detection of IUPAC and IUPAC-like Chemical Names. *Bioinformatics*. Proceedings of the International Conference Intelligent Systems for Molecular Biology (ISMB). accepted.

C. Kolářik, M. Hofmann-Apitius, M. Zimmermann, and J. Fluck. 2007. Identification of new drug classification terms in textual resources. *Bioinformatics*, 23(13):i264–i272.

J. D. Lafferty, A. McCallum, and F. C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proc. of the Eighteenth International Conference on Machine Learning (ICML 2001)*, pages 282–289. Morgan Kaufmann Publishers.

A. D. McNaught and A. Wilkinson. 1997. *Compendium of Chemical Terminology – The Gold Book*. Blackwell Science.

M. Narayanaswamy, K. E. Ravikumar, and K. Vijay-Shanker. 2003. A biological named entity recognizer. In *Proc. of the Pacific Symposium on Biocomputing*, pages 427–438.

S. J. Nelson, T. Powell, and B. L. Humphreys. 2002. The unified medical language system (umls) project. In Allen Kent and Carolyn M. Hall, editors, *Encyclopedia of Library and Information Science*, pages 369–378. Marcel Dekker, Inc, New York.

D. Rebholz-Schuhmann, H. Kirsch, M. Arregui, S. Gaudan, M. Riethoven, and P. Stoehr. 2007. EBIMed – text crunching to gather facts for proteins from medline. *Bioinformatics*, 23:237–244.

B. Sun, Q. Tan, P. Mitra, and C. L. Giles. 2007. Extraction and search of chemical formulae in text documents on the web. In *Proc. of the International World Wide Web Conference*, pages 251–260, May.

P. Tomasulo. 2002. ChemIDplus – super source for chemical and drug information. *Med Ref Serv Q*, 21(1):53–59.

D. L. Wheeler, T. Barrett, D. A. Benson, et al. 2008. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 36:D13 – D21, January.

D. S. Wishart, D. Tzur, C. Knox, R. Eisner, et al. 2007. HMDB: the human metabolome database. *Nucleic Acids Res*, 35(Database issue):D521–D526, Jan.

D. S. Wishart, C. Knox, A. C. Guo, et al. 2008. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res*, 36(Database issue):D901–D906, Jan.