Detection of IUPAC and IUPAC-like chemical names

Roman Klinger1,*, Corinna Kolářík1,2, Juliane Fluck1, Martin Hofmann-Apitius1,2 and Christoph M. Friedrich1

1Fraunhofer Institute Algorithms and Scientific Computing (SCAI), Department of Bioinformatics, Schloß Birlinghoven, 53754 Sankt Augustin and 2Bonn-Aachen International Center for Information Technology (B-IT), Department of Applied Life Science Informatics, Dahlmannstrasse 2, D-53113 Bonn, Germany

ABSTRACT

Motivation: Chemical compounds like small signal molecules or other biological active chemical substances are an important entity class in life science publications and patents. Several representations and nomenclatures for chemicals like SMILES, InChI, IUPAC or trivial names exist. Only SMILES and InChI names allow a direct structure search, but in biomedical texts trivial names and IUPAC like names are used more frequent. While trivial names can be found with a dictionary-based approach and in such a way mapped to their corresponding structures, it is not possible to enumerate all IUPAC names. In this work, we present a new machine learning approach based on conditional random fields (CRF) to find mentions of IUPAC and IUPAC-like names in scientific text as well as its evaluation and the conversion rate with available name-to-structure tools.

Results: We present an IUPAC name recognizer with an F1 measure of 85.6% on a MEDLINE corpus. The evaluation of different CRF orders and offset conjunction orders demonstrates the importance of these parameters. An evaluation of hand-selected patent sections containing large enumerations and terms with mixed nomenclature shows a good performance on these cases (F1 measure 81.5%). Remaining recognition problems are to detect correct borders of the typically long terms, especially when occurring in parentheses or enumerations. We demonstrate the scalability of our implementation by providing results from a full MEDLINE run.

Availability: We plan to publish the corpora, annotation guideline as well as the conditional random field model as a UIMA component.

Contact: roman.klinger@scai.fraunhofer.de

1 INTRODUCTION AND RELATED WORK

Finding relevant information is one of the most important challenges in our time. In particular in life science and chemical research a huge amount of new publications, research reports and patents is produced every year. For users of huge text corpora like MEDLINE, document categorization, ranking and finding entity-related information is an important help in their daily research and work life. The automated identification of entities of interest in text in a domain and their mapping to database entries strongly improves information retrieval, information extraction and information aggregation. For such tasks tools have been successfully developed in the last decades especially for finding mentions of proteins and genes. Those provide basic methods to extract e.g. protein-protein relations. For that field of interest, the BioCreative competition (Hirschman et al., 2007) provides an evaluation of state-of-the-art techniques on publicly available data sources. Dictionary-based systems allow a direct mapping of the recognized entities to reference objects (e.g. EntrezGene identifiers for genes). An inherent drawback of such approaches, however, is the dependence on the quality and completeness of the dictionary and the methods of the underlying algorithm to recognize spelling variants and to resolve ambiguous names.

Dictionary independent methods (rule-based systems as well as machine learning-based system) are well suited to find names where no comprehensive dictionary is available. An example for a rule-based approach is to find mentions matching a given set of regular expressions. Machine learning approaches are based on an annotated training set from which statistical information can be obtained about the inherent dependencies in the data. This extracted information can then be applied on unseen data to classify word tokens, i.e. to label parts of text with different classes. The best approaches in the BioCreative sub task of gene mention recognition have an F1 measure between 86% and 87% (Wilbur et al., 2007). From 21 submissions, 11 use conditional random fields (CRF), a machine learning method based on undirected graphical models (Bishop, 2006) which leads to competitive results, in our hands 86.35% F1 measure (Klinger et al., 2007b).

Finding mentions of chemical compounds in text is of interest for several reasons. An annotation of the entities enables a search engine to return documents containing elements of this entity class (semantic search), e.g. together with a disease. This can be helpful to find relations e.g. to adverse reactions or diseases. Mapping the found entities to corresponding structures leads to the possibility to search for relations between different chemicals. This enables a chemist to search for similar structures or substructures and combine the knowledge in the text with knowledge from databases or to integrate other tools handling chemical information (e.g. solubility or mass calculation).

Chemical names can be distinguished into different classes: to deal with complex structures, different methods of nomenclature are used, e.g. mentions of the sum formula or names according to the Simplified Molecular Input Line Entry Specification (SMILES; Weininger, 1988) or the successor of SMILES, the IUPAC International Chemical Identifier (InChI). Because of a limited readability of such specifications for humans, trivial 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.

Trivial names can be searched for with a dictionary-based approach and directly mapped to the corresponding structure...
at the same time. For example, the dictionary-based named entity recognition system ProMiner (Hanisch et al., 2005) uses a DrugBank1 dictionary for the recognition of drug names in MEDLINE abstracts (Kolářík et al., 2007). Other systems use the drug dictionary from MedlinePlus2 (e.g. EbiMed; Rebholz-Schuhmann et al., 2007) for the recognition of drug names.

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. Several systems address that problem regarding the identification of chemical entities with a variety of approaches. Narayanaswamy et al. (2003) describe a manually developed set of rules relying upon lexical information, linguistic constraints of the English language and contextual information for the detection of several entity classes. The reason for choosing this approach is stated as the lack of an annotated corpus. The evaluation was done on a small hand-selected corpus containing 55 MEDLINE abstracts selected by searching for acetylates, acetylated and acetylation. They found 158 chemical names from which 22 were ambiguous and classified into different classes and 13 chemical parts with two ambiguous ones. The \( F_1 \) measure for the first is 90.86% (93.15% precision, 86.08% recall). The latter has an \( F_1 \) measure of 91.67% (100% precision, 84.62% recall).

Similarly, (Kemp and Lynch, 1998) identify chemical names in patent texts with handcrafted rules using dictionaries with chemical name fragments. They claim to identify 97.4% from 14 855 specific chemical names in 70 patent descriptions taken from documents from the IPC class CO 7D. The false positive rate is reported to be 4.2%.

The concept described by Anstein et al. (2006) for which the preconditions are described by Reyle (2006) uses a grammar for the analysis of fully specified (e.g. 7-hydroxyhexan-2-one), trivial (e.g. benzene) and semi-systematic (e.g. benzene-1,3,5-triacetic acid) as well as underspecified (e.g. deoxyxugar) compound names. The advantage of that approach is that the grammatical analysis can be used as a basis for a conversion to the chemical structure. A possible problem is the difficulty to recognize names not following the specification to a certain degree as well as the completeness and maintenance of a changing standard.

A molecular similarity search is used by (Rhodes et al., 2007) to enable a user to ‘search for related Intellectual Property’ in US patents based on a specified drawn molecule. They report to find 3 623 248 unique chemical structures from 4 375 036 US patents. The absolute numbers of found patents for the top 25 drugs listed by Humana (2005) are given.

The program developed by (Sun et al., 2007) focuses on finding sum formulas like \( CH_2(CH_2)OH \) in text using support vector machines (Scholkopf and Smola, 2002) and CRFs (Lafferty et al., 2001).

In the work of Corbett et al. (2007), first-order Hidden Markov Models (Rabiner, 1989) implemented in the toolkit LingPipe3 are combined with other methods for the identification of chemical entities. The program finds e.g. structural classes, atoms and elements, fragments, trivial names, SMILES and InChI as well as IUPAC names. They give an inter-annotator \( F_1 \) measure of 93% for chemical names on their annotated corpus. The performance is evaluated on different corpora, recall rates are between 69.1% and 80.8% and precision rates between 64.1% and 75.3% (Corbett and Murray-Rust, 2006). A separate evaluation of the included named entity recognition modules from the toolkit LingPipe results in an \( F_1 \) measure of 74% (Corbett et al., 2007). To our knowledge, their implementation—the open source program OSCaR4 (Open Source Chemistry Analysis Routines; Corbett, 2007)—is the only software available to the academic community.

We prefer to have a method identifying IUPAC and IUPAC-like names only and to have additional approaches to recognize other chemical name classes (e.g. brand names or elements): IUPAC and IUPAC-like names can be recognized based on their morphological structure with higher performance than with methods based on dictionaries (Kolářík et al., 2008). We therefore introduce a system for the recognition of IUPAC and IUPAC-like names while trivial names are found with a dictionary approach not described here. These IUPAC-like terms do not only include correct IUPAC names but also names not following the nomenclature strictly. This enables a higher recall regarding mentioned chemicals, which is important for document retrieval purposes.

In addition to the correct recognition of IUPAC and IUPAC-like names, the aim is to transform these names using name-to-structure converters to allow the usage of chemical tools on the extracted data. Therefore enumerations have to be detected with all parts while modifying tokens (e.g. substitutes, analogs) have to be tagged separately (cf. Fig. 1). We use a CRF approach and present our development of a training corpus as well as our experiences regarding inter-annotator agreement in comparison to the work of Corbett et al. (2007). Next to the training corpus we describe test corpora especially on MEDLINE and on US patents.

Additionally, to the presentation of the results an exhaustive analysis of the influences of different CRF orders and offset conjunctions is shown and the impact of the different feature sets on the results are evaluated and discussed.

2 METHODS

2.1 Overview

We apply a CRF to build a model for finding IUPAC and IUPAC-related MODIFIER entities. A training corpus has been annotated by two independent annotators and the inter-annotator agreement is discussed in Section 2.3. The model selection is performed by bootstrapping (Efron and Tibshirani, 1993) and evaluated on two independent test corpora, one consisting of sampled abstracts from MEDLINE, the other one on hand selected paragraphs from bio-chemical patents. We analyze the use of name-to-structure converters as a basis of a possible normalization, a mapping of the found entities to a unique structure.

2.2 Entity types

The entities in which we focus are IUPAC and MODIFIER mentions. As described, chemical entities in general are named following different nomenclatures which are also combined by the authors of biomedical texts. Only concentrating on correct IUPAC terms is not
sufficient, so we define a IUPAC entity to be a chemical substance mentioned in a IUPAC-like manner. Additionally to correct IUPAC names, it includes IUPAC names in which is abbreviated, fragments and group names. In Figure 1 an example abstract from MEDLINE with annotations of the two entity classes is shown. Next to full names like ‘1,2,3,4-tetrahydronaphthalene-1-carboxylic acid’ or ‘4-[(3-chlorophenyl)amino][methyl]-6,7-dihydroxychromen-2-one’, fragments, e.g. in enumerations, are tagged separately like ‘acidine-4-’ and ‘phenamine-1-carboxamide’ in ‘…both the acidine-4- and ‘phenamine-1-carboxamide’ series…’ or ‘3-[bis(4-fluoro- or 4-chlorophenyl)methoxy] tropane…’.

The alternative to the separate way of annotating in enumerations would have been an annotation including the connecting word (in that example ‘or’). This is not meaningful because parts of names are sometimes divided by long text passages. With our kind of annotation, a possible enumeration resolution of the found parts in the text is prepared.

The MODIFIER entity describes similarities to a mentioned substance like ‘[IUPAC-entity] analogues’ or ‘[IUPAC-entity] modiﬁers’ or ‘3-substituted-[IUPAC-entity]-entity’.

### 2.3 Corpus generation and inter-annotator agreement

There main corpora are generated for building the model and evaluating our approach following a developed annotation guideline. A training corpus consisting of MEDLINE abstracts (abbreviated as Traina) is a test corpus containing MEDLINE abstracts (Testa) and a test corpus made up of parts of patents (Testp).

The training corpus is built in two steps. First, a preliminary corpus (named as Trainp) is built in the same manner as described by Friedrich et al. (2006). For that, in the BioCreative training corpus (Hirschman et al., 2006), the gene and protein names are replaced by randomly selected correct entities. Additionally to correct IUPAC names from PubChem (NCBI, 2007). This leads to an artificial corpus of patents (Trainp) and a test corpus made up of parts of patents (Testp).

The text is prepared by an independent person. The first measure for the IUPAC entity is relatively low with 78% (in contrast to 95% claimed by Corbett et al. (2007)). One reason for the difference in comparison to Corbett and his colleagues is our differentiation of the IUPAC entity to other chemical mentions, which is not always easy to decide while all chemical mentions in the corpus generated by Corbett are combined in one entity. Another reason are the different experience levels of our annotators: while the first annotator collaborated on the development of the annotation guideline and annotated several corpora, the second annotator based his annotations directly on the provided guideline.

In conclusion, the training corpus both annotations are combined by an independent person. The second measure between the resulting training corpus (Trainp) and the first-annotated corpus is 94%.

### 2.4 Conditional random fields

CRFs (Lafferty et al., 2001; McDonald and Pereira, 2005) are a family of probabilistic, undirected graphical models for computing the probability $P(y|x)$ of a possible label sequence $y = y_1, y_2, ..., y_n$ given the input sequence $x$. In the context of named entity recognition this observation sequence $x$ corresponds to the tokenized text. This is the sequence of tokens which is defined by a process called tokenization—splitting the text at white space, punctuation marks and parentheses. A straightforward idea for the tokenization of IUPAC names is to keep the whole name together to use the shear length for their identification. That is not possible because of leading and successive brackets and other symbols as well as often used wrong white spaces (based on e.g. converted line breaks). So we use a fine tokenization, also splitting at all number-letter changes in the text.

The label sequence is encoded in a label alphabet similar to $\mathcal{L} = \{I-IUPAC, O-B\text{-entity}, \ldots\}$. In Section 2.2.2, an example for an observation sequence with a label sequence is depicted in Figure 2.

A CRF in general is an undirected probabilistic graphical model

$$P(y|x) = \frac{1}{Z(x)} \prod_{i=1}^{n} \Psi(y_i, x)$$

where $\Psi$ are the different factors given through an independency graph like in Figure 3 (Kschischang et al., 2001). These factor functions combine

---

Fig. 1. Example abstract with tagged entities (PMID 9240357; Guzikowski et al., 1997). IUPAC entities are depicted in red while MODIFIER entities are shown in blue.
Table 1. Features used in the CRF

<table>
<thead>
<tr>
<th>Name</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static morphological features</td>
<td>Reg.Ex.</td>
</tr>
<tr>
<td>All Caps</td>
<td>([A-Z]+)</td>
</tr>
<tr>
<td>Real Number</td>
<td>((-0.9){[.-\s]}(0-9){.})]</td>
</tr>
<tr>
<td>Is Quote</td>
<td>({\ldots})</td>
</tr>
<tr>
<td>Is Slash</td>
<td>({\ldots})</td>
</tr>
<tr>
<td>Autom. generated features</td>
<td>Automatic generation of a feature for every token: match that prefix or suffix (length (k))</td>
</tr>
<tr>
<td>Bag-Of-Words</td>
<td>Automatic generation of a feature for every token: match that token</td>
</tr>
<tr>
<td>Spaces</td>
<td>White space preceding token</td>
</tr>
<tr>
<td>Spaces_left</td>
<td>White space following token</td>
</tr>
<tr>
<td>Lists</td>
<td>Prefixes and suffixes (length 4) of intermediate or last words generated from IUPAC names</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 2. Example for observation and label sequence for the text snippet: ‘…of cyclohepta-1,3-diene …’ after tokenization.

Fig. 3. First-order linear-chain CRF.

Fig. 4. Second-order linear-chain CRF.

2.5 Conversion of IUPAC names to structures

To normalize the found names, one solution is to convert them to a structure representation. Several tools have been developed for that task.
The feature set of the system with the best performance consists of last or intermediate tokens from IUPAC names is considered. From the set of static morphological features, All Caps, Real Number, Is Dash, Is Slash and Is Quote are used. The Spaces features to determine if the token is preceded or succeeded by white space is also included. Many other features, mainly from the field of gene and protein recognition were also tested, e.g. mapping the token to regular expressions representing greek letters, combinations of alpha-numeric symbols, natural numbers, etc. For lists of different features see McDonald et al. (2004), McDonald and Pereira (2005), Settles (2005) and Klüger et al. (2007a,b).

To evaluate the impact of the different features we omit one from the best feature set in several experiments (Fig. 5) and train models only with small feature sets (depicted in Fig. 6). The automatically generated features Bag-Of-Words and Autom. Prefixes/Suffixes have the highest impact on the performance together with the Spaces feature. Especially the last one is essential to obtain good results with impacts between 6.5% (CRF 2, OC 2) and 13.64% (CRF 1, OC 0). In contrast, the static morphological features and the Prefix/Suffix lists bring nearly no loss omitting them and low results when used as the only feature. Nevertheless, together with the feature Spaces, the results are surprisingly high (70% F1 measure). Interesting is that using only Autom. Prefixes/suffixes or Bag-Of-Words together with the Spaces feature and CRF order 3 and offset conjunction order 2 results in an F1 measure of 76.03% or 79.31%, respectively.

We evaluate different configurations of the features with different orders of offset conjunction (adding context in the form of features of the last p and next p tokens, where p is the order of the offset conjunction) as well as the order of the CRF, which includes information from the last q labels (q is the order of the CRF). The results of some of the features for different orders of offset conjunction and CRF can also be seen in Figure 5. The importance of the different features is nearly the same for all the different orders. The divergence in the results is high for different feature sets, but it is also very important to have the context information provided by the offset conjunction. The best F1 measure can be obtained with an offset conjunction of order 2 and a CRF of order 2 or 3. The difference between a CRF without an offset conjunction (i.e. order 0) to a CRF with order 1 offset conjunction are much higher than between order 1 and order 2 offset conjunctions. The increase of the order comes along with a high increase in the number of weights λi [compare to Equation (2) and (3)]. We have (for the CRF with order 3) 119,884 weights without an offset conjunction, 315,377 with the offset conjunction order of 2. This corresponds to the training and tagging durations depicted in Figure 7.

Inspecting the tagging errors, we find that especially boundary errors at the end or at the beginning of the name are more frequent for a lower order of the offset conjunction. Other taggings that can be correctly identified with an offset conjunction order of 2 are formulations like ‘… through the 7- or 12-methylene carbon with …’ where the high context information is necessary to classify ‘7-’ correctly. A similar example is ‘… 2,3-substituted …’ with a tagging of ‘2.3’ as IUPAC with an offset conjunction of 1 but a correct result with an offset conjunction of 2.

3.2 Evaluation of the named entity recognition

Using the best configuration identified in the previous section, the resulting model is evaluated on the sampled MEDLINE test
Detection of IUPAC-like chemical names

Fig. 5. Results on the training data TrainM with 30-fold bootstrapping with different feature sets, different orders $q$ of the CRF (given as CRF $q$ above) and different orders $p$ of the offset conjunction (given as OC $p$). The best results were obtained with the feature set presented in Table 1. For more details see Sections 2.4.1 and 3.1.

Fig. 6. Results on the training data TrainM with some very small feature sets on different orders of the CRF and offset conjunctions. In contrast to Figure 5, where variations of a larger feature set are shown, the importance of features is presented in the context of very small feature sets. (Note the different scales between Figs 5 and 6.)

corpus TestM. In Figure 7 different orders of the CRF and the offset conjunction together with tagging and training durations are depicted. Similar to the results estimated with bootstrapping on the training corpus TrainM, highest performance is obtained with the most context information included by a CRF order of 3 and an offset conjunction order of 2. The $F_1$ measure for IUPAC entities is 85.6% with a precision of 86.5% and a recall of 84.8%. The MODIFIER entities are found with an $F_1$ measure of 84.6% (91.7% precision and 78.6% recall). Higher orders have not been applied because of prohibitive training durations. However, it can be seen that our best result is obtained at the expense of a high training time and, what is more important, on a higher tagging time of 307 s then other configurations of the CRF. For tagging a higher amount of data like the full MEDLINE database one could prefer to use a faster configuration like the one with order 2 and offset conjunction of 1 which only takes 215 s for tagging the test corpus. The $F_1$ measure for IUPAC entities is lower with 77.7%, but the recall is nearly on the same level with 82.1% (MODIFIER: 55% precision, 78.6% recall). It can be concluded, that there is a trade-off between tagging time and performance, so it depends on the application which configuration should be preferred. The analysis of the errors show frequent problems in the recognition of short chemical names. On the one hand, chemical names are recognized by the system which are not specified as IUPAC-like. On the other hand, short names, similar to trivial names, specified as IUPAC-like by the annotators are most frequently unrecognized by the system. Nearly 50% of the other false positive errors are boundary errors. In addition, names morphological similar to IUPAC names like enzymes (e.g. ‘2-phospho-D-glyceratehydro-lyase’ or ‘pyruvate O2-phosphotransferase’) are detected as false positive matches.

Applying the best system trained on the MEDLINE training corpus, TrainM, for tagging the patent test corpus, TestP, shows a decrease in $F_1$ measure in comparison to the MEDLINE test corpus TestM due to the bias of hand selecting difficult paragraphs instead of sampling from a set of sentences or text snippets. We get an $F_1$ measure of 81.5% with a precision of 77.2% and a recall of 86.4%.

3.3 Annotation of full MEDLINE

We performed a run of the best CRF model on the full MEDLINE with 16 848 632 MEDLINE article entries (version as of July 13, 2007). In these entries, we have 8 975 073 abstracts. We tag titles and
From these results we cannot generalize that it is harder to find IUPAC names in patents than in abstracts, although 2.2 × 10^9 tokens, in which there are 1,715,263 IUPAC entities in 875,102 MEDLINE database entries. The tagging is performed on a computer cluster using 48 machines with two Opteron AMD double core processors with 2.6 GHz and 8 GB main memory on each machine in 76.65 h (3.19 days). The operating system is Suse Linux Enterprise Server 9 (x86-64) with the Sun N1 Grid Engine 6.

From the found IUPAC entities, only 142,181 could be transformed to a structure (16.24%). The top 15 found terms from MEDLINE are shown in Table 2, the top 5 of the converted structures in Table 3 together with the most often used terms which lead to the normalization. To get an upper bound of convertible IUPAC information by MEDLINE are shown in Table 2, the top 5 of the converted structures.

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>16811</td>
<td>N-methyl-D-aspartate</td>
</tr>
<tr>
<td>15275</td>
<td>5-hydroxytryptamine</td>
</tr>
<tr>
<td>11690</td>
<td>5-fluorouracil</td>
</tr>
<tr>
<td>9001</td>
<td>6-hydroxydopamine</td>
</tr>
<tr>
<td>7023</td>
<td>glucose-6-phosphate</td>
</tr>
<tr>
<td>6685</td>
<td>N-ethylmaleimide</td>
</tr>
<tr>
<td>5932</td>
<td>N-acetylcysteine</td>
</tr>
<tr>
<td>5178</td>
<td>12-O-tetradecanoylphorbol-13-acetate</td>
</tr>
<tr>
<td>5032</td>
<td>methyl</td>
</tr>
<tr>
<td>4742</td>
<td>N-acetylglucosamine</td>
</tr>
<tr>
<td>4311</td>
<td>benzylpenicillin</td>
</tr>
<tr>
<td>4164</td>
<td>3-methylcholanthrene</td>
</tr>
<tr>
<td>3991</td>
<td>4-aminopyridine</td>
</tr>
<tr>
<td>3931</td>
<td>2,3,7,8-tetrachlorodibenzo-p-dioxin</td>
</tr>
<tr>
<td>3979</td>
<td>5-hydroxyindocepacetic acid</td>
</tr>
</tbody>
</table>

Fig. 7. Results on the sampled MEDLINE corpus Test.med with different orders \( q \) of the CRF (given as order \( q \)) and orders \( p \) of the offset conjunction (given as \( OCP \)). The upper chart shows the \( F_1 \) measures for the different configurations, the lower one the tagging and training durations.

4 SUMMARY AND DISCUSSION

In this article, we present an approach of finding IUPAC-like terms in text with CRF. We demonstrate that our IUPAC recognizer identifies entities with an \( F_1 \) measure of 86.5% on a sampled independent test corpus built from MEDLINE. This corpus gives an estimation for all available abstracts from that database. These results show that restricting the recognition to the special class of IUPAC-like terms instead of all mentions of chemical names as focused on in Oscar3 (Corbett et al., 2007) increases the performance.

Using a tweaked corpus with correct IUPAC names shows that incorporating only complete IUPAC names in the training corpus is not sufficient. Obviously, the challenge is to recognize fragments and parts of IUPAC names. An error analysis of the final system on MEDLINE shows that boundary problems and the recognition of shorter chemical names lead to the main performance loss. This may be founded in ambiguities in the training data regarding these names and should be considered in a further extension of the training corpus. Preliminary results on an extended annotation of short names show an increase of precision to 91.4% (Kolářik et al., 2008).

When the IUPAC recognizer is applied to a hand-sampled patent corpus containing long enumerations and mixtures of different chemical nomenclatures the drop in performance is unexpectedly low with an \( F_1 \) measure of 81.5%. Apparently, the loss of \( F_1 \) measure in comparison to the MEDLINE corpus is due to a loss of precision rather than recall. Typical problems are finding the right borders of the chemical names in enumerations. From these results we cannot generalize that it is harder to find IUPAC names in patents than in abstracts.

In the feature evaluation we show that automatically generated features like Bag-of-Words and Autom. Prefixes/Suffixes together with Space information are the most important features influencing the performance of the system. The usage of combinations of these features alone e.g. Space together with prefixes and suffixes result in an \( F_1 \) measure of 76.03%. In contrast, the static morphological features which are usually very important for a good generalization (together with other morphological features), in particular on the entity class of genes and proteins (Klinger et al., 2007b) do not have such a high impact here. Remarkably, the Prefix/Suffix lists (used for generalization purposes) appear to be of very low importance indicated by nearly no loss when left out. When used as the only feature, no positive result can be obtained. However, when combined with the feature Spaces, the results are surprisingly high (70.71% \( F_1 \) measure).

Higher orders of the CRF in combination with high order offset conjunctions lead to the best results observed (\( F_1 \) measure 85.6%) on the MEDLINE test corpus with an CRF order 3 and an offset conjunction of 2. On this corpus also the direct dependency of training and labeling durations to the orders of CRF and offset conjunction are shown (cf. Fig. 7).

Despite of the analysis given here for IUPAC entities, the open question remains, in which cases a representation of context information on the labels should be preferred in comparison to a representation of context information in the text, in form of features, used here by incorporating offset conjunction. To our knowledge, no deeper analysis is published about that topic so far. Our own experiments with different orders of CRF and offset conjunctions lead to the best results observed (\( F_1 \) measure 85.6%) on the MEDLINE test corpus with an CRF order 3 and an offset conjunction of 2. On this corpus also the direct dependency of training and labeling durations to the orders of CRF and offset conjunction are shown (cf. Fig. 7).

Despite of the analysis given here for IUPAC entities, the open question remains, in which cases a representation of context information on the labels should be preferred in comparison to a representation of context information in the text, in form of features, used here by incorporating offset conjunction. To our knowledge, no deeper analysis is published about that topic so far. Our own experiments with different orders of CRF and offset conjunctions lead to the best results observed (\( F_1 \) measure 85.6%) on the MEDLINE test corpus with an CRF order 3 and an offset conjunction of 2. On this corpus also the direct dependency of training and labeling durations to the orders of CRF and offset conjunction are shown (cf. Fig. 7).
Table 3. Top 5 found converted structures (applying Ormus and CDK, drawn with Marvin [ChemAxon, 2007]) with their frequency and the frequency of occurrences of the top 3 terms which lead to the SMILES string

<table>
<thead>
<tr>
<th>Frequency</th>
<th>SMILES and example names</th>
<th>Termfreq</th>
</tr>
</thead>
<tbody>
<tr>
<td>4099</td>
<td>NC1=CC=NCC=N1 C1=NC</td>
<td>3991</td>
</tr>
<tr>
<td></td>
<td>4-amino-pyridine</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>4-Aminopyridine</td>
<td>36</td>
</tr>
<tr>
<td>3770</td>
<td>OCCS</td>
<td>3696</td>
</tr>
<tr>
<td></td>
<td>2-mercaptoethanol</td>
<td>47</td>
</tr>
<tr>
<td></td>
<td>2-Mercaptoethanol</td>
<td>19</td>
</tr>
<tr>
<td>2799</td>
<td>C1=NC2=NC=N(C2)=N1(N1)s</td>
<td>2766</td>
</tr>
<tr>
<td></td>
<td>6-mercaptopurine</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td>6-Mercaptopurine</td>
<td>7</td>
</tr>
<tr>
<td>2607</td>
<td>C1=NC=CC=C1=CC=C=C=C2</td>
<td>2416</td>
</tr>
<tr>
<td></td>
<td>1-methyl-4-phenyl-1,2,3,6-tetrahydropyridine</td>
<td>170</td>
</tr>
<tr>
<td></td>
<td>methyl-4-phenyl-1,2,3,6-tetrahydropyridine</td>
<td>10</td>
</tr>
<tr>
<td>2457</td>
<td>OCC=NC=CC=CC=1[N]1=O]=O)=O-</td>
<td>2383</td>
</tr>
<tr>
<td></td>
<td>2,4-dinitrophenol</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td>2,4-Dinitrophenol</td>
<td>11</td>
</tr>
</tbody>
</table>

Even from correct names provided by the NCBI in the database PubChem, only 30% can be converted. Unfortunately, it is not allowed to evaluate the conversion rate of the commercial tools for academic applications. Difficulties in the name to structure conversion are mixed nomenclatures and formally incorrect IUPAC and chemical names instead of correct nomenclature. We conclude that name-to-structure conversion in its current form seems to be a persistent scientific challenge.

In the future it is necessary to combine different existing tools and programs to be developed which find mentions of trivial names, formulas, IUPAC names, InChI, SMILES, group names, etc. and determine the intersection in their results and enable them also to find combined terms (e.g. in which part of the names follows the nomenclature of SMILES and another part follows IUPAC nomenclature). For that purpose, a representative, comprehensive test corpus including all these entities has to be developed.

Another goal is to combine the knowledge in drawn structures with the information in text, using the results provided by tools like chemOCR [Algorri et al., 2007; Zimmermann et al., 2005] which converts drawn structures into computer interpretable representations.

ACKNOWLEDGEMENTS

We thank all our co-workers, especially Erfan Younesi, Harsha Gurulingappa for annotation and Katrin Tomane (University of Jena) and Theo Mevissen for fruitful discussions.

Funding: This work has been partially funded by the MPG-FhG Machine Learning Collaboration (http://flg.fml.tuebingen.mpg.de/).

Conflict of Interest: none declared.

REFERENCES


Guzikowski,A.P. et al. (1997) Synthesis of racemic 6,7,8-triethylidihydro-1-benzazepine-2,5-diones as antagonists of N-methyl-D-aspartate (NMDA) and α-amino-3-hydroxy-5-methylisoxazole-4-propionic acid (AMPA) receptors. J. Med. Chem. 40, 2424–2429.


