Identifying and Classifying Terms in the Life Sciences: The Case of Chemical Terminology

Stefanie Anstein*, **Gerhard Kremer**†, **Uwe Reyle**†

*EML Research
Schlosswolfsbrunnenweg 33
D–69118 Heidelberg
Stefanie.Anstein@eml-r.villa-bosch.de
† Institute for Computational Linguistics
Azenbergstr. 12
D–70174 Stuttgart
{Gerhard.Kremer,Uwe.Reyle}@ims.uni-stuttgart.de

Abstract

Facing the huge amount of textual and terminological data in the life sciences, we present a theoretical basis for the linguistic analysis of chemical terms. Starting with organic compound names, we conduct a morpho-semantic deconstruction into morphemes and yield a semantic representation of the terms' functional and structural properties. These semantic representations imply both the molecular structure of the named molecules and their class membership. A crucial feature of this analysis, which distinguishes it from all similar existing systems, is its ability to deal with terms that do not fully specify a structure as well as terms for generic classes of chemical compounds. Such 'underspecified' terms occur very frequently in scientific literature. Our approach will serve for the support of manual database curation and as a basis for text processing applications.

1. Introduction

Because of the vast and growing amount of data in biochemical literature, natural language processing has become crucial for scientific progress. The current bottleneck thereby consists in term identification, i. e. the recognition and classification of terms as well as their mapping to a reference ID (Krauthammer and Nenadić, 2004).

The need for automatic term identification is ubiquitous, e. g. for the population and the curation of biological databases. It is part of programs that support annotators and curators of databases and resources in providing and maintaining high quality of the enormous amount of data they have to deal with. Prominent among these applications are data integration, verification and validation. Term identification is also an essential and indespensable part of a variety of computer programs within the areas of Information Retrieval, Information Extraction and Question Answering. Despite the availability of numerous terminological resources, the process of term identification is difficult (i) because there is no guarantee that these resources actually contain the entity a given term refers to, and (ii) because authors make extensive use of synonyms, alternative names and their morpho-syntactic variations.

To support the database curation as well as the information extraction task (for an overview see Šarić (2005)), we have developed a system that understands organic chemical terminology. The system, as described in Anstein and Kremer (2005), analyses fully specified (e. g. *7-hydroxyheptan-2-one*), trivial (e. g. *benzene*) and semi-systematic (e. g. *benzene-1,3,5-triacetic acid*) as well as underspecified (e. g. *deoxysugar*) compound names. It generates rich semantic representations of their molecular structure which can, e. g., be transferred to machine-readable SMILES strings[1] and determines the chemical classes the compound belongs to. Its depth of analysis and its ability to cope with underspecification and class names distinguishes it from existing systems like ChemFinder[2], PubChem[3], the 'Chemical Entity Relationships Skill Cartridge'[4] or the tools of the 'Murray-Rust Research Group' [5]. The system will, on the one hand, be used to automatically detect synonymous entries as well as errors and inconsistencies in or between databases and, on the other hand, it may serve as a basis for information extraction methods.

The role that organic chemical terminology plays in many areas of biology, in particular in cell-biology, is pivotal. Chemical nomenclature principles (e. g. IUPAC Commission on Nomenclature of Organic Chemistry (1993)) and naming conventions are not only used within the core of chemistry itself, but are among others also present in the naming of enzymes, proteins and other biochemically relevant molecules. In addition, very often some of these principles are used by authors to enrich verbs describing chemical reactions, like *phosphorylate*, with prefixes that make these reaction descriptions more precise, yielding e. g. *di-*, *5[']-*, *tyrosine-*, or *dephosphorylate*.

This paper focusses on the theoretical background of the approach to linguistically analyse chemical terminology and to provide a deep semantic representation. We will elaborate on the background and the theory of our system.

[1] SMILES: Simplified Molecular Input Line Entry System, see `http://www.daylight.com/dayhtml/smiles`

[2] `http://chemfinder.cambridgesoft.com`

[3] `http://pubchem.ncbi.nlm.nih.gov`

[4] `http://www.temis.com/index.php?id=25&selt=14&lg=en`

[5] described at `http://www.dspace.cam.ac.uk/handle/1810/740`

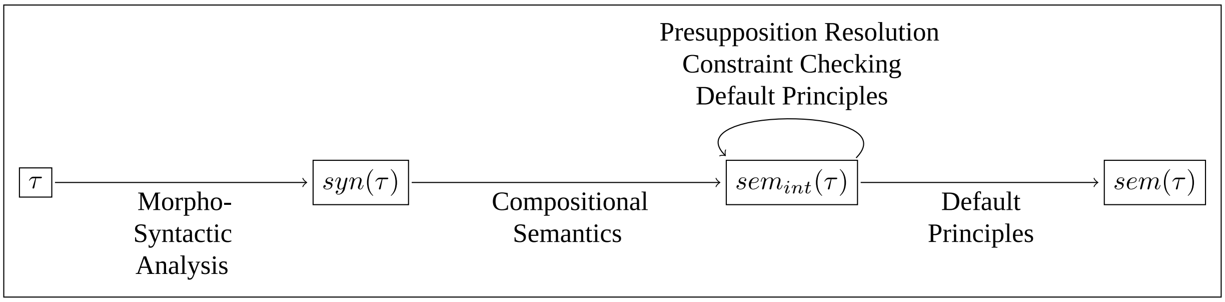

Figure 1: Sequence of analysis steps

2. Approach

Our system calculates the structural and functional aspects of molecules denoted by organic chemistry terms.

Various nomenclatures, e. g. IUPAC, specify how names for chemical compounds have to be generated from their molecular structure. These nomenclature systems serve as the basis for our symbolic, rule-based morphological analysis. The morphemes collected in a lexicon are associated with semantic expressions. These are combined compositionally to yield a semantic representation of the chemical compound name. After applying presupposition resolution and modifying the intermediate semantic representation according to default principles (illustrated in figure 1), both the identification and the classification modules use this final semantic representation to produce their results.

2.1. Morpho-Syntax

The morpho-syntactic analysis of linguistic entities is performed by a bottom-up, left-to-right parser designed with rules according to IUPAC nomenclature. The following is an example for a IUPAC nomenclature rule:

"*R-1. 2.1 Substitutive Operation:*

The substitutive operation involves the exchange of one or more hydrogen atoms for another atom or group. This process is expressed by a prefix or suffix denoting the atom or group being introduced (see R-3.2 and R-4 for lists of prefixes and suffixes)."

Such rules then lead to grammar rules allowing affixes being attached to base morphemes (in the following: parent terms), which describe the molecular skeleton structure. In general, a systematic name may consist of a parent term, prefixes and a suffix. Affixes may consist of a list of locants determining the place of operation, a multiplier and a morpheme determining the kind of operation.

An extensive lexicon is used, where variants of morphemes have separate lexicon entries. A chemical term is thus not preprocessed and pre-tokenised into its morphemes (cf. Gerstenberger (2001)), but directly analysed by the grammar. Multiple word expressions such as *pentanoic acid* are also covered by allowing space characters in certain grammar rules.

2.2. Semantics

In parallel to parsing, the semantic information contained in the lexicon is combined in a way that the resulting semantic expression represents parent term, prefixes and suffixes as determined in the syntactic structure.

The algorithm is strictly modular in the sense that it allows each meaning-bearing term component to make its own, separate contribution to the semantic representation of the term as a whole.

Following Reyle (2005), the semantic representation of a term τ is reached in several stages, as depicted in figure 1. The interpretation algorithm starts with a morpho-syntactic analysis as described in section 2.1. yielding $syn(\tau)$, and then constructs in a bottom-up fashion an intermediate semantic representation $sem_{int}(\tau)$ by inserting the semantic contributions of τ's morphemes into $syn(\tau)$ and calculating the meanings of more complex constituent parts of τ along the principles of dynamic semantics (Groenendijk and Stokhof, 1991). The intermediate representation $sem_{int}(\tau)$ then undergoes procedures that possibly enrich it and check its consistency. Enrichment is achieved by resolution of presuppositions (see Kamp et al. (2004)) and by application of default rules used in nomenclature operations. Presupposition-carrying morphemes are in particular locants, and default rules govern element and binding types as well as the presence of hydrogen atoms. Consistency of intermediate representations is checked wrt a valence model. The resulting presupposition-free representation will then, again by application of default principles, be transformed into the final representation, $sem(\tau)$.

The semantic representation yielded looks as follows: A predicate *compd* contains three arguments, viz the semantic description of (i) the parent part of the compound name, (ii) the name's prefix(es) and (iii) the name's suffix as in '*compd(ParentTerm,Prefixes,Suffix)*'.

2.3. Identification via SMILES strings

Term identification requires the representation of the molecular structure of a given compound name in a machine-readable way. A SMILES string serves this purpose as it represents even a complex molecular structure by help of a line-based notation system and various tools exist to process them.

We create, from the semantics of a name, an intermediate structure representation coded in a predicate-argument term. This term includes all information on atoms, bonds, etc. that can be derived from the name according to our semantics construction. In the next step, this intermediate representation is transferred into a corresponding SMILES string.

For underspecified compound names such as *hydroxyheptan-2-one*, where the locant of the *hydroxy-*

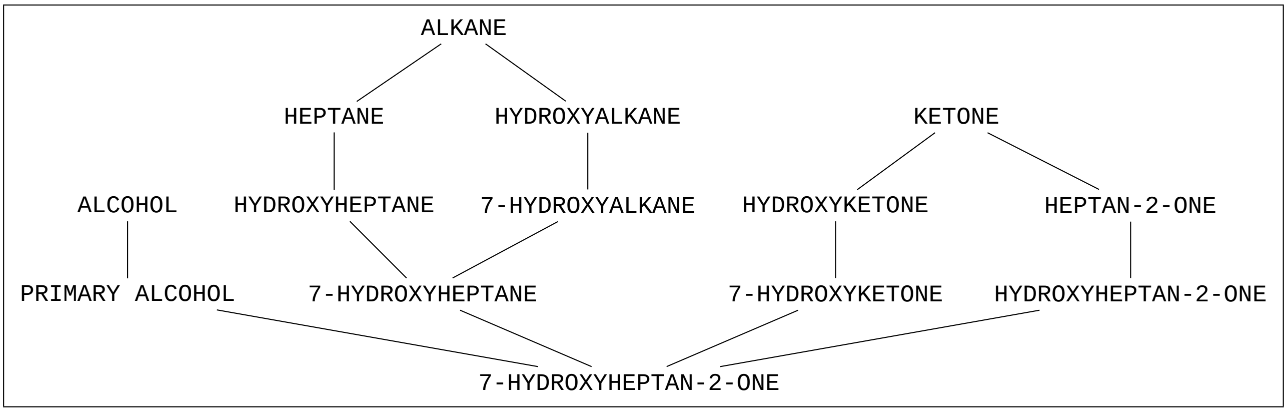

Figure 2: Class hierarchy for *7-hydroxyheptan-2-one*

prefix is missing, partial SMILES strings are generated. In certain cases a list of possible resolutions can be offered, e. g. yielding the predicate-argument term *underspecified(CC(=O)CCCCC , [{1,3,4,5,6,7}-hydroxy])*. Thus, even if not all information about a molecule is given in its name, some information about it can nevertheless be used in subsequent processing of the results.

2.4. Classification

The classes a compound belongs to are also calculated on the basis of the name's morphemes and semantic representation. In some cases, a direct morpheme-class mapping can be done; in other cases, more complex methods of class assignment have to be applied. The latter occurs, e. g., if affix morphemes 'interact' with parent morphemes, i. e. if an affix describes a change in the skeleton chain which influences also the compound's class.

Intermediate classes (as shown in figure 2) become crucial for automatic intelligent text processing, e. g. if an article about specific compounds only mentions one of their superclasses in the title. An example for a publication containing such an intermediate class (viz *Hydroxyketone*) is the paper "Ozonolysis of Alkenes and Study of Reactions of Polyfunctional Compounds: LXIII. A New Procedure for Direct Reduction of 1-Methylcycloalkene Ozonolysis Products to Hydroxyketones"[6].

Such a hierarchy can be generated automatically on the basis of our analysis in that all possible intermediate classes are calculated. By abstracting step by step, a compound can thus be classified from its most specific superclass to the most general one. This information is also crucial for building up complex ontologies as knowledge bases for scientists, where such a hierarchy would be part of.

2.5. Beyond taxonomic relations

Classification of chemical compounds according to functional and chemical properties induces a refinement of the reactions they participate in (see Wittig et al. (2004)). Very often, however, either not all subclasses of a compound class may be substrates of a reaction, or the output of a reaction depends on the subclasses. In both cases it is neccessary to have access to knowledge about the subclasses that goes beyond mere classification, in particular knowledge about the chemical structure of compounds. Take, for example, the enzymatic reaction *protein N-phosphohistidine + sugar = protein histidine + sugar phosphate*[7]. This equation states that some *sugar* is phosphorylated by transfer of the phosphate group in *protein N-phosphohistidine*. However, it is left underspecified which sugars may be substrates of the reaction and where they are phosphorylated. Only in the comment lines added to the reaction description in the enzyme database we learn that "Aldohexoses, and their glycosides and alditols, are phosphorylated on O-6, whereas fructose and sorbose are phosphorylated on O-1". A desideratum would thus be to replace the general reaction description by a set of more specific ones, which in this case would have the form *protein N-phosphohistidine + aldohexose = protein histidine + aldohexose 6-phosphate*, or *protein N-phosphohistidine + fructose = protein histidine + fructose 1-phosphate*. It is important to note, however, that neither the form of the general reaction description nor the form of the more specific ones suite for computer applications without a formal analysis of what the terms they contain denote. Even if a (rather sophisticated) classification of chemical compounds covered these terms by saying that, e. g., an aldohexose 6-phosphate is (a subclass of) an aldohexose phosphate, the differentiation actually made in the comment would not be accounted for, because it is not so much the subclass relation that is important here, but the fact that for one subclass phosphorylation takes place at O-6 and for the other subclass at O-1. It follows that the ontological knowledge needed to deal with such reaction specifications must go beyond a classification of compounds in terms of being subclasses of each other. It must be able to interleave knowledge about aspects of the molecular level of compounds with knowledge about the reactions they may be substrates of.

[6] Ishmuratov, G. Y.; Kharisov, R. Y.; Yakovleva, M. P.; Botsman, O. V.; Muslukhov, R. R.; Tolstikov, G. A. 2001. *Russian Journal of Organic Chemistry*, 37(1):37-39.

[7] see enzymatic classification No. EC 2.7.1.69, (IUBMB, 1992)

3. Results

Our deep analysis approach yields as its first outcome the detailed representation of a chemical term's semantics. SMILES strings are then provided on the basis of these semantics to identify a term's molecular structure. Additionally, a list of classes a chemical compound belongs to is calculated, also according to its semantic representation.

These results are important for the support of manual database population, integration and curation as well as for text processing tasks. Additionally, the system may be applied to analyse and formalise chemical reaction descriptions possibly containing underspecified terms and class names. The expressive and deductive power of the semantic representation language for molecular structures and reactions outranges the SMILES representation language and any of its extensions.

4. Conclusion

Our approach to understanding organic compound names is to be seen as a basis for further enhancements regarding more complex chemical terminology. It can also be transferred to other domain terminology. For a valuable implementation of the theory presented, a high-quality lexicon is crucial, especially also for the treatment of trivial and semi-systematic compound names. In the identification and classification task, underspecified names are an important issue as they appear often in scientific literature. It is only with a deep linguistic approach such as the one presented here that underspecified terminology can be handled.

The population, integration and curation of high-quality biochemical databases as well as intelligent text processing methods for the handling of the existing huge amount of data are highly dependent on terminology treatment and, especially, understanding. This is where our linguistic approach provides valuable support of life science research.

5. Acknowledgements

Stefanie Anstein is funded by the Klaus Tschira Foundation gGmbH, Heidelberg (`http://www.kts.villa-bosch.de`).

6. References

Stefanie Anstein and Gerhard Kremer. 2005. Analysing Names of Organic Chemical Compounds – From Morpho-Semantics to SMILES Strings and Classes. Master's thesis, IMS, University of Stuttgart. Web version available at `http://www.ims.uni-stuttgart.de/lehre/studentenarbeiten/fertig/Diplomarbeit_Anstein_Kremer.pdf`.

Ciprian V. Gerstenberger. 2001. Semantische Analyse von Namen Organischer Verbindungen oder Was Bedeutet 3,3'-Ureylen-dibenzamidin? Master's thesis, University of Stuttgart.

Jeroen Groenendijk and Martin Stokhof. 1991. Dynamic predicate logic. *Linguistics and Philosophy*, 14:39–100.

IUBMB. 1992. *Enzyme Nomenclature*. Academic Press, San Diego, California.

IUPAC Commission on Nomenclature of Organic Chemistry. 1993. *A Guide to IUPAC Nomenclature of Organic Compounds (Recommendations 1993)*. Blackwell Scientific publications. Web version retrieved November 2005, from `http://www.acdlabs.com/iupac/nomenclature`.

Hans Kamp, Josef van Genabith, and Uwe Reyle. 2004. Discourse Representation Theory. In Dov Gabbay and Franz Günthner, editors, *Handbook of Philosophical Logic*. Kluwer, Dordrecht.

Michael Krauthammer and Goran Nenadić. 2004. Term Identification in the Biomedical Literature. *Journal of Biomedical Informatics (Special Issue on Named Entity Recognition in Biomedicine)*, 37(6):512–526.

Uwe Reyle. 2005. Understanding Chemical Terminology. *Terminology*. Retrieved November 2005, from `ftp://ftp.ims.uni-stuttgart.de/pub/papers/reyle/terminology.pdf`.

Jasmin Šarić. 2005. *Information Extraction for Biology*. Ph.D. thesis, University of Stuttgart.

Ulrike Wittig, Andreas Weidemann, Renate Kania, Christian Peiss, and Isabel Rojas. 2004. Classification of Chemical Compounds to Support Complex Queries in a Pathway Database. *J. Comp. Funct. Genom.*, 5(2):156–162, March.