Pyridines, pyridine and pyridine rings: disambiguating chemical named entities

Peter Corbett*, Colin Batchelor†, Ann Copestake‡

*Unilever Centre for Molecular Science Informatics, Cambridge University Chemical Laboratory,
Lensfield Road, Cambridge UK CB2 1EW
†Royal Society of Chemistry, Thomas Graham House, Cambridge UK CB4 0WF
‡University of Cambridge Computer Laboratory, 15 JJ Thompson Avenue, Cambridge UK CB3 0FD
*ptc24@cam.ac.uk, †batchelorc@rsc.org, ‡aac10@cl.cam.ac.uk

Abstract

In this paper we investigate the manual subclassification of chemical named entities into subtypes representing whole compounds, parts of compounds and classes of compounds. We present a set of detailed annotation guidelines, and demonstrate their reproducibility by performing an inter-annotator agreement study on a set of 42 chemistry papers. The accuracy and κ for the annotating the subtypes of the majority named entity type were 86.0% and 0.784 respectively, indicating that consistent manual annotation of these phenomena is possible. Finally, we present a simple system that can make these judgments with accuracy of 67.4% and κ of 0.470.

1. Introduction

Pyridines are chemical compounds that contain a pyridine ring, and the simplest pyridine is pyridine itself.[1] Here we see an ambiguity whereby a chemical name can be used to refer to a specific compound, a class of compounds, or a part of compound. This is not unique to pyridine; it is a form of regular polysemy which applies to almost all names of chemical compounds. This is a significant problem for chemical named entity recognition,[2] and is not addressed by the annotation guidelines for existing corpora that mark up chemical names, such as the BioIE P450 corpus (Kulick et al., 2004), Genia (Kim et al., 2003) or our own chemical-focused corpus and annotation scheme (Corbett et al., 2007). The aim of our current work is to build on our existing annotation scheme for chemistry named entities in a way which captures this systematic ambiguity.[3]

The compound/class of compounds polysemy is an example of autohypernymy (or autosuperordination (Cruse, 2004)), and the (class of) compound/part of compound polysemy is an example of automeronymy, similar to certain sorts of metonymy. Examples of similar phenomena can be found in other domains. For example, in animals it is common to have a word that refers to both an animal of unspecified gender and a specific gender of that animal (for example, "cow" can mean a female bovine or any bovine), and in some cases a word can refer to different taxonomic levels (for example, "cat" can refer both to the domestic cat *Felis silvestris catus* or to the Felidae in general). Food and drink provide further examples of this; one can distinguish between "Merlot" as a mass noun, denoting the wine, and "a Merlot" as a count noun, denoting a particular sort of Merlot wine, and also "Merlot grapes". For people, one may say "a Kennedy" to mean a member of the Kennedy family, and "Kennedy" to mean some specific member of that family (usually JFK)—thus it may be said that Kennedy was a Kennedy. Nevertheless the regular polysemy of chemical names is particularly interesting, not least because the distinctions between the different senses can be mapped onto well-defined notions in the domain of chemical structure.

In many cases the compound, class-of-compounds and part-of-compound senses are grammatically marked in a regular fashion: exact compounds tend to occur as singular bare noun phrases, as mass (or maybe proper) nouns; classes of compounds tend to occur in the plural or with a determiner, as count nouns; parts of compounds tend to occur in noun–noun compounds, with a head noun such as "ring", "chain", "group", "moiety", "substituent" or "subunit".[4] However these patterns are frequently violated. For example, a compound with a pyridine ring may be described as being "pyridine-containing" or "bearing a pyridine". These sense distinctions are defined in terms of chemical structure, and may be highly correlated with grammatical cues, but they are not fundamentally grammatical distinctions.

It should be noted that the compound/class of compound distinction cuts across distinctions between kind-referring and substance- or specimen-referring uses. For example, "pyridine has a boiling point of 115 °C" can be considered as kind-referring,[5] whereas "the laboratory had to be evacuated because pyridine had been spilled on the floor" refers to a specific specimen of pyridine. However, there is

[1] A pyridine ring is five carbon atoms and one nitrogen atom in a ring, linked by alternating single and double bonds as in benzene. In pyridine (C_5H_5N) each carbon atom is bonded to a hydrogen atom. In other pyridines one or more of these hydrogens are substituted by a different atom or group.

[2] For a review of chemical named entity recognition, see Banville (2006) and the introduction to Corbett *et al.* (2007).

[3] We do not consider non-systematic ambiguity in this paper, which for chemistry named entities chiefly arises from acronyms.

[4] In "pyridine ring", the part-of-compound sense applies to the whole noun–noun compound. The sense of the word "pyridine" itself within the compound is harder to define: it may be argued that it is ambiguous between the exact-compound sense ("the ring found in pyridine"), the class-of-compounds sense ("the ring that defines pyridines") and possibly the part-of-compound sense (in a redundant construction similar to "pine tree"). However, if we need to annotate individual words, it makes sense from a practical point of view to annotate the word with the part-of-compound sense.

[5] Carlson *et al.* (1995) make the same point about "gold is a precious metal", with the footnote: "We say 'can be considered' because we do not wish to prejudge the final semantic analysis to be given."

a common factor in both these cases: the complete chemical structure of pyridine, C_5H_5N. This factor also applies in the cases of pyridine as part of a mixture of solvents, pyridine molecules diffusing through a membrane, the abstract pyridine molecule and a computer simulation of pyridine. By contrast, this factor is not present in class-of-compound uses. For example, "pyridines such as pentafluoropyridine", where the pyridine ring is present, but the five hydrogen atoms are not. It is even possible to use the class-of-compound use to refer to a specimen: "Yesterday I synthesised pentafluoropyridine and tetrafluoropyrrole. Today I spilled the pyridine but I still have the pyrrole."

Regular polysemies also exist for other types of chemical named entities. Some are traditionally regarded as the names of classes of compounds, for example "alkene" or "ester". These do not have an exact-compound sense but they do possess a part-of-compound sense as well as their class-of-compound sense. Underspecified names (Reyle, 2006) such as "dimethylpyridine" (which does not say where the methyl groups are attached to the pyridine ring) also have this property. Other names, such as "methyl", are most commonly used for parts of compounds; however these may also be used to refer to specific entities (in compounds such as "methyl radical" or "methyl ion") or to classes of compounds (for example "methyl compounds").

The names of chemical elements display an extended form of this regular polysemy. For example, "carbon" has a bulk-substance sense similar to the exact-compound sense for "pyridine" and an atom sense similar to the part-compound sense "pyridine ring". The class sense also exists in the form of "carbon compounds". Finally there is arguably a fourth sense that is unique to elements. One may, for example, talk of "atmospheric carbon". This is not a synonym for soot particles, but refers to all of the carbon atoms in atmospheric carbon dioxide, carbon monoxide, methane and other carbon compounds as if they could some single substance. Note that this is not synonymous with "atmospheric carbon compounds": one tonne of atmospheric carbon dioxide accounts for only 0.27 tonnes of atmospheric carbon. This sense is unusual, as chemically it is like the part-of-compound sense, but it typically has the grammatical markings of the exact-compound sense, and is typically substance-referring, rather than kind-referring.

Other forms of chemical nomenclature, such as chemical formulae and acronyms, may also participate in these systems of exact-compound, part-of-compound and class-of-compound senses. A few of these entities may be regarded as monosemous. For example, "$-CH_3$" unambiguously refers to a methyl group, a part of a compound.

We present an approach to disambiguation which involves both (i) determining which senses are available for a given named entity, and then (ii) disambiguating the entity based on its context. We collapse these into a single classification task, where a single list of senses, or "subtypes",[6] is used as the set of possible classifications for all chemical names. This formulation as a classification problem follows Markert and Nissim (2002) and SemEval 2007 Task #8 (Markert and Nissim, 2007) in their approach to metonymy.

These subtype distinctions could be used for a number of purposes. Subtype information could be included in patterns for information extraction. For example, in the extraction of "A is-a B" relations, the precision could be increased by only allowing relations where A is `EXACT` or `CLASS`, and B is `CLASS`. Subtypes could also be useful in an information retrieval context. For example a user could seach for "pyridine" as `EXACT` only, and get only mentions to the specific compound, and not to other pyridines.

Another use for subtypes is the assignment of identifiers to named entities. The chemical ontology ChEBI (Degtyarenko et al., 2008) contains entries for specific compounds, classes of compounds and parts of compounds. These may have the same or closely related names. For example there is an entry for "thiol group" (CHEBI:29917), corresponding to the part-of-compound subtype, which `is_part_of` "thiols" (CHEBI:29256), corresponding to the class-of-compounds subtype.

Furthermore, with chemicals it is often possible to concisely annotate a chemical name with the structure of the chemical compound that it represents. There are a variety of useful formats for storing the structures of specific compounds, some of which (such as SMILES and InChI) have useful advantages in terms of conciseness and canonicalisation, whereas there are fewer formats for storing classes of compounds, and to the best of our knowledge none of these are canonicalisable. Thus, a system of subtypes could act in a manner analogous to the type systems used in compilers, specifying what sort and format of data is needed to represent a particular entity.

In this paper we briefly mention related work in the gene/protein domain (section 2), describe an annotation scheme for disambiguating these regular polysemies of chemical named entities (section 3) and then describe an interannotator agreement experiment (section 4). We illustrate the task with some examples drawn from the corpus (section 5). Finally we present some simple systems for automated annotation of these subtypes (section 6) and suggest some directions for further work (section 7).

2. Related Work

We know of no other detailed work on this issue of exact/class/part distinctions of chemical names. However there are a few publications that touch on related issues for gene and protein named entities.

Vlachos and Gasperin (2006) divide noun phrases that contain gene names into *gene mentions* and *other mentions*. It has been shown that these distinctions can be made with 78.6% accuracy using SVMs and syntactic parsing (Korkontzelos et al., 2007). Gasperin *et al.* (2007) then classify their gene NPs as `gene`, `product`, `subtype`, `part-of`, `supertype` or `variant`.

In the Genia ontology (Kim et al., 2003), which provides the type system for the Genia corpus, there are subtypes of nucleic acids and proteins, dealing with both parts and classes of these biomacromolecules. However, there is no such subtyping for their more chemical named entity types, we are unaware of any inter-annotator agreement studies

[6] We refer to subtypes rather than senses, and in general speak of subclassification or subtyping rather than disambiguation, due to the presence of occasional monosemous cases.

for these distinctions, and much of the named-entity recognition work based on the Genia corpus ignores these distinctions entirely, as it uses the simplified version of the corpus used in the JNLPBA evaluations.

3. Annotation Guidelines

The work presented in this paper is based upon the named entity annotation scheme of Corbett et al. (2007) (henceforth, "the named entity guidelines"), which has five classes of named entity. These guidelines were applied to a corpus of 42 chemistry papers, three each from fourteen journals covering most of chemistry (henceforth "the corpus"), and it was shown that the guidelines could be applied with 93% inter-annotator agreement. The majority class, CM ("chemical"), accounted for 94.1% of the named entities in the, and includes specific chemicals, parts of chemicals and some classes of chemicals[7] with no disambiguation between them. The second most common class, RN ("reaction"), was intended to mainly cover words that denoted (a subset of) chemical reactions; however, it also included instances of these words that were used to describe chemicals, or denote the bulk movement of chemicals.

The subtypes annotation task consists of taking a corpus that has previously been annotated for named entities according to the named entity guidelines, and assigning one and only one subtype to each entity that belongs to the type CM ("chemical"), RN ("reaction"), CJ ("chemical adjective") or ASE ("enzyme"). There are no subtypes for the type CPR ("chemical prefix"). Each type has its own list of available subtypes, and the subtype OTHER was available for exceptionally difficult cases (it is expected that the use of OTHER will occur less than once per paper).

We have developed a set of annotation guidelines (henceforth "the guidelines"), specifying the subtypes available for each named entity type, and providing advice on difficult distinctions. These were developed in an iterative process, where a proposed set of guidelines were used in an informal inter-annotator agreement study on a batch of test papers, and experience from that study was used to inform the refinement of the guidelines. None of the 42 papers in the corpus was used or referred to in this process.

The subtypes annotation applies to the named entities themselves, and not in general to the noun phrase that contains them. This makes annotation easier and allows the scheme to remain agnostic about linguistic issues surrounding the structure of noun phrases. However in many cases the guidelines allow head words in noun-noun compounds to be used as cues; for example in "pyridine ring", "ring" is a cue that "pyridine" should be given the subtype PART.

The predominant named entity class in the corpus is CM (chemical), encompassing 95% of the entities. We divide this into six subtypes: three major subtypes, EXACT (for specific compounds), CLASS (for classes of compounds) and PART (parts of compounds, and classes of those parts), to deal with a large majority of the named entities, and three minor subtypes, SPECIES (corresponding to the fourth subtype identified for elements in the introduction, the sense of "atmospheric carbon"), and SURFACE (for surfaces) and POLYMER (for polymers), to deal with special cases that did not fit well with the major classes.

3.1. EXACT, CLASS and PART

We distinguish EXACT from PART according to whether the author was talking about an entity as some free item that had an existence of its own right, rather than making a judgement according to what sort of chemical bonds there were inside and outside of the entity.

Our distinction between EXACT and CLASS is different from the issue of genericity and specificity, for example as discussed by Herbelot and Copestake (2008). Our CLASS subtype deals with situations where the named entity itself does not specify a specific compound. CLASS applies both to terms denoting entire classes of chemicals and to members of that class that are not specified by the named entity itself, including anaphoric and deictic uses. So "pyridines" would typically be CLASS, as would "pyridine" in "The Hantzsch pyridine synthesis is the formation of a pyridine from an aldehyde, a ketoester and a nitrogen donor" and in "the pyridine **6**" (where **6** is a reference to a structural diagram denoting a specific pyridine). EXACT is used for the pyridine that is called "pyridine", as in "dissolved in pyridine." Genericity annotations in the style of Herbelot and Copestake could conceivably be applied to entities of type CLASS as an additional stage of processing.

3.2. SPECIES, SURFACE and POLYMER

The minor subtypes were introduced to cover particular difficulties caused by the major subtypes not entirely fitting the domain. These subtype distinctions take precedence over the distinctions between major subtypes: for example, in "sodium halide surface" and "sodium chloride surface", "sodium chloride" and "sodium halide" are annotated as SURFACE, even though "sodium chloride" would normally be EXACT and "sodium halide" CLASS.

The subtype SPECIES (which is often easier to annotate than to concisely explain or come up with a good name for), exemplified by "atmospheric carbon", arises from the fact that the number of atoms of a given element is usually conserved. In essence, SPECIES is considering atoms as part of a bulk sample, rather than as part of the structure of a compound, which would warrant the use of PART. There are a number of contexts that are particularly associated with SPECIES, for example environmental or metabolic processes associated with a particular element, toxic metals such as lead, mercury or polonium, and elemental analysis techniques such as ICP. The name SPECIES derives from the fact that in these cases the atoms of the element are often said to belong to different chemical species.

The subtype SURFACE was added because surfaces introduce a confounding part-of relation. A surface is a part of a specimen of bulk material but not a part of a chemical structure in the way that for example a pyridine ring might be. Also, there are forms of notation for specific types of surface that are included within some named entities. For example "Ag(111)" represents a specific surface of a crys-

[7] The classes of chemicals had to be those that could be defined at least partially in terms of structure (*e.g.* "pyridines") and/or elemental composition (*e.g.* "hydrocarbons"): classes that were purely based on origin (*e.g.* "natural product") or activity (*e.g.* "antioxidant") were excluded.

tal of silver with a specific arrangement of silver atoms. This arrangement of atoms, different from the arrangement in Ag(110), gives the Ag(111) surface different properties from Ag(110), even though the two surfaces may be different faces of the same specimen of silver.

Polymers are particularly difficult, hence the POLYMER subtype. Many polymer samples consist of a mixture of polymer molecules of varying sizes and shapes, so we could imagine several different part-of and class-of subtypes for polymers. As polymers are moderately rare in general chemistry corpora it makes sense to group all of these together as POLYMER to avoid too many complications.

3.3. RN, CJ and ASE

The types other than CM accounted for only 5.9% of the named entities in the corpus between them, and so less attention was devoted to them. Briefly:

The type RN ("reaction") was divided into three subtypes: REACT (actual reactions), DESC (descriptions of compounds), and MOVE (bulk movements of compounds). For example, the word "chlorinated" would be REACT in "benzene was chlorinated to give chlorobenzene", DESC in "chlorobenzene is a chlorinated compound" and MOVE in "We chlorinated the swimming pool".

The type CJ ('chemical adjective') had subtypes EXACT, CLASS and PART, by analogy with CM, and also ACID, SOLUTION and RECEPTOR. The names of many acids are of the form "<something>ic acid". Sometimes the "<something>ic" word gets detached from its "acid", for example one could talk about "citric acidity". In this case "citric" would get the subtype ACID. SOLUTION is used for words such as "aqueous" or "ethanolic" when they are used to describe solutions, and RECEPTOR is used for words such as "nicotinic" and "adrenergic" that are used to describe types of receptors.

The type ASE, which covers words derived from chemical names that end in -ase, has two subtypes, PROTEIN, where the word, for example "demethylase", refers to a protein, and ACTIVITY, where the word refers to the ability of a protein to perform the function of, say, a demethylase - i.e. to catalyse demethylation reactions. Although ASE was rare (0.7% of entities) in the corpus, we observe that the ACTIVITY subtype is common in some subdomains. For example, in the cytochrome P450 literature, it is common to refer to the enzymatic demethylation of aminopyrine as "aminopyrine demethylase activity", even though at least five different enzymes have this activity.

4. Inter-annotator Agreement

To test the annotation guidelines, we performed an inter-annotator agreement study. We used the corpus of 42 chemistry papers of Corbett *et al.* (2007). These papers had been selected from those published by the Royal Society of Chemistry in January 2004, randomly selecting 3 full papers or short papers from each non-review journal. Those papers had then been annotated for named entities. The named entities in that work had been assigned to five categories, CM, RN, CJ, CPR and ASE. Only 14 papers had been annotated by all three annotators, whereas all 42 had been annotated by their Subject A, who is our annotator A and the first author of this paper (Corbett). They did not produce an adjudicated annotation.

We subclassify Subject A's CM, RN, CJ and ASE annotations with the subtype scheme described above. There were two subjects, annotators A and B, who are the first two authors of this paper (Corbett and Batchelor) and the authors of the guidelines.[8] They are both trained chemists and were Subjects A and B in Corbett et al. (2007). We annotated the papers with a custom-made software tool tool that presented the annotators with a drop-down menu which allowed them to select exactly one subtype for each of the named entities.

During annotation the subject were allowed to refer to the annotation guidelines, to reference sources, to their domain knowledge as chemists, and to the original chemistry papers (including the figures). They were not allowed to confer with anyone over the annotation, nor to refer to texts annotated during development of the guidelines.

We use two metrics to assess interannotator agreement; accuracy and κ (kappa). Accuracy is simply the proportion of named entities for which both annotators gave the same type. The κ metric is a more sophisticated measure which factors out random agreement. We use Cohen's κ (Cohen, 1960), where the distribution of categories is calculated independently for each annotator.

4.1. Results and Discussion

Class	N	n	Accuracy	κ (k = 2)
CM	6865	6	86.0%	0.784
RN	288	3	95.5%	0.828
CJ	60	6	75.0%	0.363
ASE	31	2	90.3%	−0.045

Table 1: Inter-annotator agreement by named entity class. N is the number of entities. n is the number of available subtypes (excluding OTHER). k is the number of annotators.

In Table 1 we can see that the κ values are acceptable (above 0.67) for both CM and RN, whereas CJ and ASE were not reproducibly annotated. However, between them, CJ and ASE accounted for only about 1% of the named entities in the corpus, and so this is not a major issue. Note that the results in this table represent four essentially independent experiments, one per named entity class.

The results for RN are very encouraging - it was easy to annotate in the original named entity task too (F = 94%). REACT was the majority subtype (84% by annotator A; 85% by B), with all but one of the remaining named entities being annotated as DESC. MOVE was not adequately tested in this exercise - a corpus from more biological domains, such as physiology and cell biology might be a better test of this subtype.

CJ proved to be problematic during the original named-entity annotation too (inter-annotator F = 56%), suggesting

[8] This may have resulted in slightly inflated scores for inter-annotator agreement, due to tacit understandings being developed during guidelines development. This has been demonstrated previously, for example Corbett *et al.* (2007)

that CJ represents a rather ill-defined collection of phenomena. Most examples of CJ (68% by annotator A; 88% by B) were of the subtype SOLUTION.

ASE was reliably (F = 96%) annotated in the original named entity task. Almost all incidences of ASE were annotated as PROTEIN, there was two cases where one annotator chose ACTIVITY, one case where the other annotator chose it, and no cases where both annotators chose ACTIVITY. The subtypes of ASE would most likely be better tested in a corpus such as the PennBioIE P450 corpus (Kulick et al., 2004).

For CM, the median accuracy was 90.7%, the minimum accuracy was 61.1%, and two papers (one containing 2 CM entities, one containing 86) were annotated with 100% accuracy. This concurs with the annotators' experiences that the subject matter and writings styles encountered in some papers presented particular difficulties.

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

Table 2: Breakdown of subtypes of CM. Columns 2 and 3 show numbers of entities found by annotator 1, columns 4 and 5 show annotator 2.

Table 2 shows a breakdown of the results for CM. 95% of the entities fell into one of the three major subtypes, with slightly less than half being EXACT. There was little need to resort to using OTHER. In general, the ease of annotating a subtype, as measured by the F score, appears to correlate quite well with how common the subtype is. This is unsurprising, as the minor subtypes were invented as a means of dealing with tricky cases.

None of the F scores seen here are as high as the 93% that was achieved for the original named entity annotation, which suggests that the subtypes task is a harder task (at least for human annotators) than named entities.

Table 3 shows the confusion matrix for CM. Intuitively one might have expected a large confusion between CLASS and PART, owing to the ambiguous use of terms to mean both functional groups and compounds possessing them, but this is not especially marked. Evidently the guidelines were largely sufficient to address this issue. EXACT/PART ambiguity appears to be a larger issue, with one annotator having a bias towards EXACT and another towards PART.

5. Examples

In these examples, named entities of type CM are underlined.

5.1. EXACT, CLASS and PART

Corbett *et al.* (2007) draw an example from the corpus which is worth repeating here:

> In addition, we have found in previous studies that the Zn^{2+}–Tris system is also capable of efficiently hydrolyzing other β-lactams, such as clavulanic acid, which is a typical mechanism-based inhibitor of active-site serine β-lactamases (clavulanic acid is also a fairly good substrate of the zinc-β-lactamase from *B. fragilis*).

In this example, "clavulanic acid" is obviously EXACT, and "β-lactams" is very strongly marked as CLASS. "Zn^{2+}–Tris" was annotated as a whole named entity, and refers to a specific complex, and so is also EXACT. More interesting is the mention of "serine", as part of the name of the enzyme family "serine β-lactamases". Here, background knowledge is required to know that the serine is mentioned as the catalytic amino acid residue, rather than as the substrate of the enzyme. As such, it is referring to a serine residue as part of a protein, and is annotated as PART; serine as a free amino acid would be annotated as CLASS.

Another, more difficult, example can be taken from the same source text:

> [...] it has been proposed that the metal ions bind to the β-lactam carboxylate group, promoting the attack of external hydroxide on the β-lactam carbonyl group.

"Carbonyl" and "carboxylate" are both clearly PART here. "Hydroxide" presumably means a hydroxide ion. The negative charge of the hydroxide ion makes it impossible to get a bottle of hydroxide ions, and hydroxide is often a part of salts such as sodium hydroxide; however, in this case it is a free species, independent of any positive ion, and so can be annotated as EXACT.

The real difficulty concerns "β-lactam", where we must disambiguate CLASS from PART (there is no such compound as "β-lactam", so EXACT is inappropriate). Here, we need background knowledge to know that β-lactam rings contain carbonyl groups, they do not contain carboxylate groups, and that the β-lactams studied in the paper do not contain carboxylate groups directly attached to the β-lactam ring system. It is also useful to know that carboxylate groups themselves contain carbonyl groups. Given this knowledge, we can deduce that the first β-lactam must be CLASS, and the second PART. In the second case, there are at least two carbonyl groups to consider: the carbonyl group in the β-lactam ring, and the carbonyl group in the carboxylate group. The authors disambiguate between these carbonyls, specifying the former of the two, with "the β-lactam carbonyl" where "β-lactam" specifies the location of the carbonyl group. However, in the case of the carboxylate group, it is neither a part of the β-lactam ring nor attached to it, so the PART reading is inappropriate and the "β-lactam" therefore must specify the whole molecule, so CLASS is the correct annotation.

In another paper, we encounter the following example, where "FA" stands for "fatty acid" and "LPS" for lipopolysaccharide (a biomacromolecule; not a chemical named entity according to the guidelines):

> After 2h of hydrolysis, 14:0 3-OH FA, the

	EXACT	CLASS	PART	SPECIES	SURFACE	POLYMER	OTHER
EXACT	**2988**	92	258	10	52	2	0
CLASS	87	**915**	90	14	8	0	0
PART	136	102	**1729**	5	4	4	2
SPECIES	27	11	28	**165**	2	0	0
SURFACE	3	0	2	0	**65**	3	0
POLYMER	3	5	10	0	0	**40**	0
OTHER	2	0	1	0	0	0	**0**

Table 3: Confusion matrix for subtypes of CM.

> only 3-OH FA in the LPS of E. coli, was detected in GC-MS analysis.

Here, "3-OH FA" refers to a fatty acid as part of a larger biomacromolecule, much like the "serine" in the first example, and is clearly PART. However, "14:0 3-OH FA" appears to be referring to the fatty acid once it has been hydrolysed from the LPS and become a free fatty acid, and thus is annotated as EXACT. This reading is further reinforced by background knowledge of GC-MS, an analytical technique that is well-suited to the detection of small molecules but which is unlikely to be useful for the detection of whole biomacromolecules.

5.2. SPECIES, SURFACE and POLYMER

> Selenized yeast contains a large number of water-soluble selenium compounds but most of the selenium is incorporated into sparingly soluble bio-molecules that are difficult to extract.

The first "selenium" is part of the phrase "selenium compounds", and therefore is annotated as CLASS, whereas the second "selenium" is annotated as SPECIES.

> In the cases of sulfate surfaces previously studied, the frictional asymmetry was detected at monatomic steps where the tilt directions of the sulfate ions were reversed.

Here the first "sulfate" takes CLASS, as the paper had mentioned minerals such as calcium sulfate and strontium sulfate; however, in this case SURFACE is required. The second case is clearly not talking about a whole surface, and so one annotator chose PART and the other EXACT.

> Both of the separations make use of resin-based anion exchange columns with quaternary ammonium fuctional *(sic)* groups. The PRP-X100 column (Separation 1) utilizes a poly(styrene–divinyl)benzene polymeric support while the IC-Pak A HR column (Separation 2) is based on a polymethacrylate resin.

The last two entities here are both POLYMER. Neither "poly(styrene–divinyl)benzene" nor "polymethacrylate" fully describe the polymeric compounds used; in both cases, the compounds incorporate some monomeric building blocks bearing ammonium groups.

6. Automated Systems

In this section we discuss some simple automated systems for assigning subtypes. These systems are intended to provide a simple measure of the difficulty of the problem, and to set a baseline for future systems. This section focuses exclusively on CM as the named entity type of interest, and uses the annotations produced by annotator A.

The simplest baseline is to annotate everything as EXACT. This achieves an accuracy of 49.5% against annotator A but a κ of zero.

To improve on this, we make use of a simple machine learning system based on a maximum-entropy classifier[9]. We evaluate the classifier using three-fold cross validation, with each fold consisting of one-third of the papers (one from each journal). We explore several possible features, considering each feature both in isolation and in combination with the other features.

For simplicity, we consider only features that are easy to obtain; we use the tokeniser described by Corbett *et al.* (2007) (and combine multi-token named entities into a single token, converting whitespace within them to underscores) and no other NLP components such as POS taggers or parsers. This is useful as we do not know of any of these which have been specifically trained for chemistry text. Furthermore, in the scenario where an NER system is run prior to parsing, as a method for unknown word identification, it seems likely that the extra information from subtype classification will be useful for a parser.

We use two types of feature, name-internal and name-external. Name-internal features often determine what subtypes are possible for a name and their probability distribution, and name-external features can be used to disambiguate these possibilities.

The two name-internal features are ("name"), the unmodified name itself which when used alone implements a first sense heuristic (see McCarthy *et al.* (2007) for a discussion of this heuristic in general WSD), backing off to the most common subtype if the name has not been seen in the training corpus, and ("suffix"), which is the last four characters of the name, or the whole name if it is shorter than that. The suffixes of chemical names are often informative; for

[9] We use the OpenNLP MaxEnt classifier, http://maxent.sourceforge.net/. If it is required to classify an instance with no features, it returns a uniform probability distribution. To fix this, and get a probability distribution representative of the training data as a whole, all training and test instances are given a single "bias" feature.

example, names ending in “yl” are likely to prefer PART (*e.g.* “methyl”), whereas names ending in “oid” are likely to be CLASS (*e.g.* “alkaloid”). A small amount of experimentation shows that four is the best number of characters to use. Finally, we can detect whether or not a name is plural (“plural”) by looking for a terminal “s”; irregular plurals are very rare in chemical names, and singular names ending in “s” (*e.g.* “chlorpyrifos”) are uncommon.

We propose two simple name-external features; the previous token (“previous”) and the next token (“next”). These tokens may include punctuation tokens. The XML format that is used for our tokens marks up citation references (which appear as numbers in superscript) - if the next or previous token is one of these, we skip it and take the next next token or previous previous token instead. Furthermore, if the next token is a hyphen, we skip over the hyphen. As such, “pyridine-based” and “pyridine based” are treated the same by our system.

The “next” feature is expected to detect many of the head words in noun–noun compounds. Often they imply PART (*e.g.* “group”, “ring”, “bond”), or spectroscopic features that derive from parts of compounds (*e.g.* “peak”, “stretch”). Some head words typically signify EXACT (*e.g.* “molecules”), whereas others typically specify CLASS (*e.g.* “compounds”). Furthermore the presence of punctuation or a common verb is likely to indicate the name is the head of its noun phrase, weighing against PART.

The previous token feature in effect combines several forms of evidence. For example, the presence of a determiner signifies that EXACT is unlikely.[10] Conversely, a preposition helps to indicate a bare noun phrase and is thus likely to constitute evidence in favour of EXACT. Some premodifiers can distinguish bulk elements from atoms of that element (*e.g.* “elemental”, “molecular”, “dry”), and some others are likely to be good indicators of SPECIES (*e.g.* “atmospheric”, “dietary”, “total”). Some premodifiers are also useful as evidence of PART, (*e.g.* “bridging”, “terminal”).

We test the system with several feature sets. The feature set “all” contains all of the features, and “none” is a feature-free setup that always selects the most common subtype. The feature set “name” only uses the name feature, “−name” uses all of the features except for the name feature. The other feature sets follow this naming scheme, except for “internal”, which is a combination of “name”, “suffix”; “plural”, and “external”, which is a combination of “next” and “previous”; and “p+p+n”, which combines “plural”, “previous” and “next”.

Table 4 shows the influence of the various features. It is clear that all five features are useful in this task, with “next” being particularly important. Interestingly, the features “plural” and “previous” can be contrasted with “name” and “suffix”, in that the former pair seem to be more important as part of a large feature collection, whereas the latter pair work quite well on their own but are less important in the combined feature set. This observation inspired the “p+p+n” feature set, which showed that removing both “name” and “suffix” had a much larger effect than removing either one of them, demonstrating a large amount of redundancy between them.

Feature set	Accuracy	κ
none	49.5%	0.0
name	56.2%	0.213
suffix	59.2%	0.303
plural	53.4%	0.114
previous	54.2%	0.208
next	61.0%	0.311
internal	60.9%	0.334
external	61.9%	0.364
p+p+n	65.6%	0.434
−name	67.3%	0.468
−suffix	67.0%	0.459
−plural	66.1%	0.447
−previous	66.7%	0.452
−next	62.0%	0.372
all	**67.4%**	**0.470**

Table 4: Automated results for CM by feature set

Subtype	Precision	Recall	*F*
EXACT	70.9%	83.4%	76.7%
CLASS	65.6%	46.1%	54.2%
PART	62.0%	57.8%	59.8%
SPECIES	53.6%	48.5%	50.9%
SURFACE	94.1%	21.9%	35.6%
POLYMER	71.4%	8.6%	15.4%

Table 5: Automated results for CM by subtype. The computer did not assign OTHER to any name.

Table 5 shows the breakdown for the “all” feature set, by subtype. The more common subtypes are more easily recognised, with recall being bad for the rare subtypes.

Overall, it is clear that the problem is at least partly tractable, but also that considerable improvements will have to be made to get good accuracy.

It is obvious that there are a number of directions in which this baseline system could be extended. As well as experimenting with different machine-learning techniques, an obvious approach is to use a parser rather than simple proximity to generate context features. It may also be possible to look at the larger context; for example, the mention of elemental analysis techniques such as ICP may indicate an increased likelihood of SPECIES. Additional name-internal features could also be useful. For example, there are a number of class terms available which can be included in systematic names, such as “alkyl”, “acyl” and “halo”, the presence of which could be used to rule out EXACT. Furthermore, the parsing and interpretation of systematic names could be helpful (Reyle, 2006). For example, “dimethylpyridine” is ambiguous; there are several different ways to put two methyl groups on a pyridine ring. This ambiguity means that EXACT is not likely to be appropriate in this case.

[10] Markert and Nissim (2005) find that the presence of a determiner is a good feature for identifying org-for-product metonymy, as is a word being plural.

7. Further Directions

The obvious next steps with this system of annotations are to experiment with greater quantities and other genres of chemical text, and to explore more sophisticated approaches to the automation of the annotation. However there are a few areas in which the annotation scheme itself could be extended.

It is clear that some of the subtypes could themselves be divided up into subsubtypes. PART is the most obvious of these, covering functional groups, rings, chains, atoms, bonds, ligands in complexes, amino acid residues in proteins and a few other systems. Furthermore it would be useful to distinguish between precisely-specified parts (e.g. "methyl") and classes of parts (e.g. "alkyl"). For CLASS it would be useful to distinguish truly generic uses from uses that mention a specific compound but not by name. For EXACT, there are cases where there is not quite enough information in the named entity itself to make a full distinction. For example, in both "sodium metal" and "sodium ion", the named entity (according to the annotation guidelines) is "sodium", and the subtype in both cases is EXACT. Resolving these cases to point to the correct entries in a database will require more information than the named entity itself and the subtype.

There are also cases where what is understood by "a specific compound" is variable. For example, lactic acid is a chiral molecule which has left- and right-handed forms, L-lactic acid and D-lactic acid. A mention of "lactic acid" in text may indicate either form, a mixture of the two, or that the author did not remember, care or know that the two forms of the compound existed.

8. Conclusion

We have identified subtypes of the chemical named entities defined by Corbett *et al.*, and have produced extensive annotation guidelines for them. We have shown that the inter-annotator agreement is acceptable for the named entity types CM and RN across the major genres of chemistry papers. Furthermore we have demonstrated a simple system for automatically making these assignments, showing that the problem is both tractable and non-trivial, and setting a baseline for future systems. These annotations will assist in the assignment of ontology identifiers to chemical named entities, and should be useful in information extraction and information retrieval systems.

The annotation guidelines are available by contacting the first author.

9. Acknowledgements

PTC thanks Peter Murray-Rust for supervision. We thank the RSC for providing the papers, and the UK eScience Programme and EPSRC (EP/C010035/1) for funding. We thank the reviewers for their helpful suggestions and regret that we did not have the time or space to address all of the issues raised.

10. References

Debra Banville. 2006. Mining chemical structural information from the drug literature. *Drug Discovery Today*, 11:35-42.

Gregory Carlson and Francis Pelletier. 1995. The Generic Book. *University of Chicago Press*

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37-46.

Peter Corbett, Colin Batchelor and Simone Teufel. 2007. Annotation of Chemical Named Entities. *BioNLP 2007: Biological, translational, and clinical language processing*, 57-64.

Alan Cruse. 2004. Meaning in Language: An Introduction to Semantics and Pragmatics. *Oxford Textbooks in Linguistics.*

Kirill Degtyarenko, Paula de Matos, Marcus Ennis, Janna Hastings, Martin Zbinden, Alan McNaught, Rafael Alcantara, Michael Darsow, Mickael Guedj and Michael Ashburner. 2008. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res*, Vol. 36, Database issue D344-D350.

Caroline Gasperin, Nikiforos Karamanis and Ruth Seal. 2007. Annotation of anaphoric relations in biomedical full-text articles using a domain-relevant scheme. *Proceedings of DAARC 2007*, 19-24.

Aurelie Herbelot and Ann Copestake. 2008. Annotating Genericity: How Do Humans Decide? (A Case Study in Ontology Extraction). *Proceedings of the Third International Conference for Linguistic Evidence.*

J.-D. Kim, T. Ohta, Y. Tateisi and J. Tsujii. 2003. GENIA corpus—a semantically annotated corpus for biotextmining. *Bioinformatics*, 19(Suppl 1):i180-i182.

Ioannis Korkontzelos, Andreas Vlachos and Ian Lewin, 2007. From gene names to actual genes. *Proceedings of BioLink, Vienna.*

Seth Kulick, Ann Bies, Mark Liberman, Mark Mandel, Ryan McDonald, Martha Palmer, Andrew Schein and Lyle Ungar. 2004. Integrated Annotation for Biomedical Information Extraction. *HLT/NAACL BioLINK workshop*, 61-68.

Katja Markert and Malvina Nissim. 2002. Metonymy resolution as a classification task. *Proceedings of the 4th Conference on Empirical Methods in Natural Language Processing (EMNLP 2002).*

Katja Markert and Malvina Nissim. 2005. Learning to buy a Renault and talk to BMW: A supervised approach to conventional metonymy. *International Workshop on Computational Semantics (IWCS2005)*

Katja Markert and Malvina Nissim. 2007. SemEval-2007 Task 08: Metonym Resolution at SemEval-2007. *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pp 36-41.

Diana McCarthy, Rob Koeling, Julie Weeds and John Carroll. 2007. Finding Predominant Word Senses in Untagged Text. *Computational Linguistics* 33 (4), pp 553-590.

Uwe Reyle. 2006. Understanding chemical terminology. *Terminology* 12:1, pp 111-126.

Andreas Vlachos and Caroline Gasperin. 2006. Bootstrapping and Evaluating Named Entity Recognition in the Biomedical Domain. *Proceedings of BioNLP in HLT-NAACL*. 138-145.