Annotation of Chemical Named Entities

Peter Corbett
Cambridge University
Chemical Laboratory
Lensfield Road
Cambridge
UK CB2 1EW
ptc24@cam.ac.uk

Colin Batchelor
Royal Society of Chemistry
Thomas Graham House
Milton Road
Cambridge
UK CB4 0WF
batchelorc@rsc.org

Simone Teufel
Natural Language and
Information Processing Group
Computer Laboratory
University of Cambridge
UK CB3 0FD
sht25@cam.ac.uk

Abstract

We describe the annotation of chemical named entities in scientific text. A set of annotation guidelines defines 5 types of named entities, and provides instructions for the resolution of special cases. A corpus of full-text chemistry papers was annotated, with an inter-annotator agreement F score of 93%. An investigation of named entity recognition using LingPipe suggests that F scores of 63% are possible without customisation, and scores of 74% are possible with the addition of custom tokenisation and the use of dictionaries.

1 Introduction

Recent efforts in applying natural language processing to natural science texts have focused on the recognition of genes and proteins in biomedical text. These large biomolecules are—mostly—conveniently described as sequences of subunits, strings written in alphabets of 4 or 20 letters. Advances in sequencing techniques have lead to a boom in genomics and proteomics, with a concomitant need for natural language processing techniques to analyse the texts in which they are discussed.

However, proteins and nucleic acids provide only a part of the biochemical picture. Smaller chemical species, which are better described atom-by-atom, play their roles too, both in terms of their interactions with large biomolecules like proteins, and in the more general biomedical context. A number of resources exist to provide chemical information to the biological community. For example, the National Center For Biotechnology Information (NCBI) has added the chemical database PubChem[1] to its collections of bioinformatics data, and the ontology ChEBI (Chemical Entities of Biological Interest) (de Matos et al., 2006) has been added to the Open Biological Ontologies (OBO) family.

Small-molecule chemistry also plays a role in biomedical natural language processing. PubMed has included abstracts from medicinal chemistry journals for a long time, and is increasingly carrying other chemistry journals too. Both the GENIA corpus (Kim et al., 2003) and the BioIE cytochrome P450 corpus (Kulick et al., 2004) come with named entity annotations that include a proportion of chemicals, and at least a few abstracts that are recognisable as chemistry abstracts.

Chemical named entity recognition enables a number of applications. Linking chemical names to chemical structures, by a mixture of database lookup and the parsing of systematic nomenclature, allows the creation of semantically enhanced articles, with benefits for readers. An example of this is shown in the Project Prospect[2] annotations by the Royal Society of Chemistry (RSC). Linking chemical NER to chemical information retrieval techniques allows corpora to be searched for chemicals with similar structures to a query molecule, or chemicals that contain a particular structural motif (Corbett and Murray-Rust, 2006). With information extraction techniques, chemicals could be linked to their properties, applications and reactions, and with traditional gene/protein NLP techniques, it could be pos-

[1] http://pubchem.ncbi.nlm.nih.gov/
[2] http://www.projectprospect.org/

BioNLP 2007: Biological, translational, and clinical language processing, pages 57–64,
Prague, June 2007.

sible to discover new links between chemical data and bioinformatics data.

A few chemical named entity recognition (Corbett and Murray-Rust, 2006; Townsend et al., 2005; Vasserman, 2004; Kemp and Lynch, 1998; Sun et al., 2007) or classification (Wilbur et al., 1999) systems have been published. A plugin for the GATE system[3] will also recognise a limited range of chemical entities. Other named entity recognition or classification systems (Narayanaswamy et al., 2003; Torii et al., 2004; Torii and Vijay-Shanker, 2002; Spasic and Ananiadou, 2004) sometimes include chemicals as well as genes, proteins and other biological entities. However, due to differences in corpora and the scope of the task, it is difficult to compare them. There has been no chemical equivalent of the JNLPBA (Kim et al., 2004) or BioCreAtIvE (Yeh et al., 2005) evaluations. Therefore, a corpus and a task definition are required.

To find an upper bound on the levels of performance that are available for the task, it is necessary to study the inter-annotator agreement for the manual annotation of the texts. In particular, it is useful to see to what extent the guidelines can be applied by those not involved in their development. Producing guidelines that enable a highly consistent annotation may raise the quality of the results of any machine-learning techniques that use training data applied to the guidelines, and producing guidelines that cover a broad range of subdomains is also important (Dingare et al., 2005).

2 Annotation Guidelines

We have prepared a set of guidelines for the annotation of the names of chemical compounds and related entities in scientific papers. These guidelines grew out of work on PubMed abstracts, and have since been developed with reference to organic chemistry journals, and later a range of journals encompassing the whole of chemistry.

Our annotation guidelines focus on the chemicals themselves; we believe that these represent the major source of rare words in chemistry papers, and are of the greatest interest to end-users. Furthermore, many chemical names are formed systematically or semi-systematically, and can be interpreted without resorting to dictionaries and databases. As well as chemical names themselves, we also consider other words or phrases that are formed from chemical names.

The various types are summarised in Table 1.

Type	Description	Example
CM	chemical compound	citric acid
RN	chemical reaction	1,3-dimethylation
CJ	chemical adjective	pyrazolic
ASE	enzyme	methylase
CPR	chemical prefix	1,3-

Table 1: Named entity types

The logic behind the classes is best explained with an example drawn from the corpus described in the next section:

> 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*).

Here, ‘clavulanic acid’ is a specific chemical compound (a CM), referred to by a trivial (unsystematic) name, and ‘β-lactams’ is a class of chemical compounds (also a CM), defined by a particular structural motif. ‘Zn^{2+}–Tris’ is another CM (a complex rather than a molecule), and despite being named in an *ad hoc* manner, the name is compositional and it is reasonably clear to a trained chemist what it is. ‘Serine’ (another CM) can be used to refer to an amino acid as a whole compound, but in this case refers to it as a part of a larger biomolecule. The word ‘hydrolyzing’ (an RN) denotes a reaction involving the chemical ‘water’. ‘β-lactamases’ (an ASE) denotes a class of enzymes that process β-lactams, and ‘zinc-β-lactamase’ (another ASE) denotes a β-lactamase that uses zinc. By our guidelines, the terms ‘mechanism-based inhibitor’ or ‘substrate’ are not annotated, as they denote a chemical role, rather than giving information about the structure or composition of the chemicals.

[3] `http://www.gate.ac.uk/`

The full guidelines occupy 31 pages (including a quick reference section), and contain 93 rules. Almost all of these have examples, and many have several examples.

A few distinctions need to be explained here. The classes RN, CJ and ASE do not include all reactions, adjectives or enzymes, but only those that entail specific chemicals or classes of chemicals—usually by being formed by the modification of a chemical name—for example, 'β-lactamases' in the example above is formed from the name of a class of chemicals. Words derived from Greek and Latin words for 'water', such as 'aqueous' and 'hydrolysis', are included when making these annotations.

The class CPR consists of prefixes, more often found in systematic chemical names, giving details of the geometry of molecules, that are attached to normal English words. For example, the chemical 1,2-diiodopentane is a 1,2-disubstituted pentane, and the '1,2-' forms the CPR in '1,2-disubstituted'. Although these contructions sometimes occur as infixes within chemical names, we have only seen these used as prefixes outside of them. We believe that identifying these prefixes will be useful in the adaptation of lexicalised parsers to chemical text.

The annotation task includes a small amount of word sense disambiguation. Although most chemical names do not have non-chemical homonyms, a few do. Chemical elements, and element symbols, give particular problems. Examples of this include 'lead', 'In' (indium), 'As' (arsenic), 'Be' (beryllium), 'No' (nobelium) and 'K' (potassium—this is confusable with Kelvin). These are only annotated when they occur in their chemical sense.

2.1 Related Work

We know of two publicly available corpora that also include chemicals in their named-entity markup. In both of these, there are significant differences to many aspects of the annotation. In general, our guidelines tend to give more importance to concepts regarding chemical structure, and less importance to biological role, than the other corpora do.

The GENIA corpus (Kim et al., 2003) includes several different classes for chemicals. Our class CM roughly corresponds to the union of GENIA's `atom`, `inorganic`, `other_organic_compound`, `nucleotide` and `amino_acid_monomer` classes, and also parts of `lipid` and `carbohydrate` (we exclude macromolecules such as lipoproteins and lipopolysaccharides). Occasionally terms that match our class RN are included as `other_name`. Our CM class also includes chemical names that occur within enzyme or other protein names (e.g. 'inosine-5′-monophosphate' in 'inosine-5′-monophosphate dehydrogenase') whereas the GENIA corpus (which allows nesting) typically does not. The GENIA corpus also sometimes includes qualifiers in terms, giving 'intracellular calcium' where we would only annotate 'calcium', and also includes some role/application terms such as 'antioxidant' and 'reactive intermediate'.

The BioIE P450 corpus (Kulick et al., 2004), by contrast, includes chemicals, proteins and other substances such as foodstuffs in a single category called 'substance'. Again, role terms such as 'inhibitor' are included, and may be merged with chemical names to make entities such as 'fentanyl metabolites' (we would only mark up 'fentanyl'). Fragments of chemicals such as 'methyl group' are not marked up; in our annotations, the 'methyl' is marked up.

The BioIE corpus was produced with extensive guidelines; in the GENIA corpus, much more was left to the judgement of the annotators. These lead to inconsistencies, such as whether to annotate 'antioxidant' (our guidelines treat this as a biological role, and do not mark it up). We are unaware of an inter-annotator agreement study for either corpus.

Both of these corpora include other classes of named entities, and additional information such as sentence boundaries.

3 Inter-annotator Agreement

3.1 Related Work

We are unaware of any studies of inter-annotator agreement with regards to chemicals. However, a few studies of gene/protein inter-annotator agreement do exist. Demetriou and Gaizauskas (2003) report an F score of 89% between two domain experts for a task involving various aspects of protein science. Morgan *et al.* (2004) report an F score of 87% between a domain expert and a systems developer for *D. melanogaster* gene names. Vlachos and Gasperin (2006) produced a revised version of the guidelines for the task, and were able to achieve an F score of 91%, and a kappa of 0.905, between a computational linguist and a domain expert.

3.2 Subjects

Three subjects took part in the study. Subject A was a chemist and the main author of the guidelines. Subject B was another chemist, highly involved in the development of the guidelines. Subject C was a PhD student with a chemistry degree. His involvement in the development of guidelines was limited to proof-reading an early version of the guidelines. C was trained by A, by being given half an hour's training, a test paper to annotate (which satisfied A that C understood the general principles of the guidelines), and a short debriefing session before being given the papers to annotate.

3.3 Materials

The study was performed on 14 papers (full papers and communications only, not review articles or other secondary publications) published by the Royal Society of Chemistry. These were taken from the journal issues from January 2004 (excluding a themed issue of one of the journals). One paper was randomly selected to represent each of the 14 journals that carried suitable papers. These 14 papers represent a diverse sample of topics, covering areas of organic, inorganic, physical, analytical and computational chemistry, and also areas where chemistry overlaps with biology, environmental science, materials and mineral science, and education.

From these papers, we collected the title, section headings, abstract and paragraphs, and discarded the rest. To maximise the value of annotator effort, we also automatically discarded the experimental sections, by looking for headers such as 'Experimental'. This policy can be justified thus: In chemistry papers, a section titled "Results and Discussion" carries enough information about the experiments performed to follow the argument of the paper, whereas the experimental section carries precise details of the protocols that are usually only of interest to people intending to replicate or adapt the experiments performed. It is increasingly common for chemistry papers not to contain an experimental section in the paper proper, but to include one in the supporting online information. Furthermore, experimental sections are often quite long and tedious to annotate, and previous studies have shown that named-entity recognition is easier on experimental sections too (Townsend et al., 2005).

A few experimental sections (or parts thereof) were not automatically detected, and instead were removed by hand.

3.4 Procedure

The papers were hand-annotated using our in-house annotation software. This software displays the text so as to preserve aspects of the style of the text such as subscripts and superscripts, and allows the annotators to freely select spans of text with character-level precision—the text was not tokenised prior to annotation. Spans were not allowed to overlap or to nest. Each selected span was assigned to exactly one of the five available classes.

During annotation the subjects were allowed to refer to the guidelines (explained in the previous section), to reference sources such as PubChem and Wikipedia, and to use their domain knowledge as chemists. They were not allowed to confer with anyone over the annotation, nor to refer to texts annotated during development of the guidelines. The training of subject C by A was completed prior to A annotating the papers involved in the exercise.

3.5 Evaluation Methodology

Inter-annotator agreement was measured pairwise, using the F score. To calculate this, all of the exact matches were found and counted, and all of the entities annotated by one annotator but not the other (and vice versa) were counted. For an exact match, the left boundary, right boundary and type of the annotation had to match entirely. Thus, if one annotator had annotated 'hexane–ethyl acetate' as a single entity, and the other had annotated it as 'hexane' and 'ethyl acetate', then that would count as three cases of disagreement and no cases of agreement. We use the F score as it is a standard measure in the domain—however, as a measure it has weaknesses which will be discussed in the next subsection.

Given the character-level nature of the annotation task, and that the papers were not tokenised, the task cannot sensibly be cast as a classification problem, and so we have not calculated any kappa scores.

Overall results were calculated using two methods. The first method was to calculate the total levels of agreement and disagreement across the whole corpus, and to calculate a total F score based on that. The second method was to calculate F scores for individual papers (removing a single paper that contained two named entities—neither of which were spotted by subject B—as an outlier), and to calculate an unweighted mean, standard deviation and 95% confidence intervals based on those scores.

3.6 Results and Discussion

Subjects	F (corpus)	F (average)	std. dev.
A–B	92.8%	92.9%±3.4%	6.2%
A–C	90.0%	91.4%±3.1%	5.7%
B–C	86.1%	87.6%±3.1%	5.7%

Table 2: Inter-annotator agreement results. ± values are 95% confidence intervals.

The results of the analysis are shown in Table 2. The whole-corpus F scores suggest that high levels of agreement (93%) are possible. This is equivalent to or better than quoted values for biomedical inter-annotator agreement. However, the poorer agreements involving C would suggest that some of this is due to some extra information being communicated during the development of the guidelines.

A closer analysis shows that this is not the case. A single paper, containing a large number of entities, is notable as a major source of disagreement between A and C, and B and C, but not A and B. Looking at the annotations themselves, the paper contained many repetitions of the difficult entity ‘Zn^{2+}–Tris’, and also of similar entities. If the offending paper is removed from consideration, the agreement between A and C exceeds the agreement between A and B.

This analysis is confirmed using the per-paper F scores. Two-tailed, pairwise t-tests (excluding the outlier paper) showed that the difference in mean F scores between the A–B and A–C agreements was not statistically significant at the 0.05 significance level; however, the differences between B–C and A–B, and between B–C and A–C were.

A breakdown of the inter-annotator agreements by type is shown in Table 3. CM and RN, at least, seem to be reliably annotated. The other classes are less easy to assess, due to their rarity, both in terms of their total occurrence in the corpus and the number of papers that contain them.

Type	F	Number
CM	93%	2751
RN	94%	79
CJ	56%	20
ASE	96%	25
CPR	77%	10

Table 3: Inter-annotator agreement, by type. F scores are corpus totals, between Subjects A and C. The number is the number of entities of that class found by Subject A.

We speculate that the poorer B–C agreement may be due to differing error rates in the annotation. In many cases, it was clear from the corpus that errors were made due to failing to spot relevant entities, or by failing to look up difficult cases in the guidelines. Although it is not possible to make a formal analysis of this, we suspect that A made fewer errors, due to a greater familiarity with the task and the guidelines. This is supported by the results, as more errors would be involved in the B–C comparison than in comparisons involving A, leading to higher levels of disagreement.

We have also examined the types of disagreements made. There were very few cases where two annotators had annotated an entity with the same start and end point, but a different type; there were 2 cases of this between A and C, and 3 cases in each of the other two comparisons. All of these were confusions between CM and CJ.

In the A–B comparison, there were 415 entities that were annotated by either A or B that did not have a corresponding exact match. 183 (44%) of those were simple cases where the two annotators did not agree as to whether the entity should be marked up or not (i.e. the other annotator had not placed any entity wholly or partially within that span). For example, some annotators failed to spot instances of ‘water’, or disagreed over whether ‘fat’ (as a synonym for ‘lipid’) was to be marked up.

The remainder of those disagreements are due to disagreements of class, of where the boundaries should be, of how many entities there should be in a given span, and combinations of the above. In all of these cases, the fact that the annotators produce at least one entity each for a given case means that disagreements of this type are penalised harshly, and therefore are given disproportionate weight. However, it is also likely that disagreements over whether to mark an entity up are more likely to represent a simple mistake than a disagreement over how to interpret the guidelines; it is easy to miss an entity that should be marked up when scanning the text.

A particularly interesting class of disagreement concerns whether a span of text should be annotated as one entity or two. For example, 'Zn^{2+}–Tris' could be marked up as a single entity, or as 'Zn^{2+}' and 'Tris'. We looked for cases where one annotator had a single entity, the left edge of which corresponded to the left edge of an entity annotated by the other annotator, and the right edge corresponded to the right edge of a different entity. We found 43 cases of this. As in each of these cases, at least three entities are involved, this pattern accounts for at least 30% of the inter-annotator disagreement. Only 17 of these cases contained whitespace—in the rest of the cases, hyphens, dashes or slashes were involved.

4 Analysis of the Corpus

To generate a larger corpus, a further two batches of papers were selected and preprocessed in the manner described for the inter-annotator agreement study and annotated by Subject A. These were combined with the annotations made by Subject A during the agreement study, to produce a corpus of 42 papers.

Type	Entities		Papers	
CM	6865	94.1%	42	100%
RN	288	4.0%	23	55%
CJ	60	0.8%	20	48%
ASE	31	0.4%	5	12%
CPR	53	0.7%	9	21%

Table 4: Occurrence of entities in the corpus, and numbers of papers containing at least one entity of a type.

From Table 4 it is clear that CM is by far the most common type of named entity in the corpus. Observation of the corpus shows that RN is common in certain genres of paper (for example organic synthesis papers), and generally absent from other genres. ASE, too, is a specialised category, and did not occur much in this corpus.

A closer examination of CM showed more than 90% of these to contain no whitespace. However, this is not to say that there are not significant numbers of multi-token entities. The difficulty of tokenising the corpus is illustrated by the fact that 1114 CM entities contained hyphens or dashes, and 388 CM entities were adjacent to hyphens or dashes in the corpus. This means that any named entity recogniser will have to have a specialised tokeniser, or be good at handling multi-token entities.

Tokenising the CM entities on whitespace and normalising their case revealed 1579 distinct words—of these, 1364 only occurred in one paper. There were 4301 occurrences of these words (out of a total of 7626). Whereas the difficulties found in gene/protein NER with complex multiword entities and polysemous words are less likely to be a problem here, the problems with tokenisation and large numbers of unknown words remain just as pressing.

As with biomedical text (Yeh et al., 2005), cases of conjunctive and disjunctive nomenclature, such as 'benzoic and thiophenic acids' and 'bromo- or chlorobenzene' exist in the corpus. However, these only accounted for 27 CM entities.

5 Named-Entity Recognition

To establish some baseline measures of performance, we applied the named-entity modules from the toolkit LingPipe,[4] which has been successfully applied to NER of *D. melanogaster* genes (e.g. by Vlachos and Gasperin (2006)). LingPipe uses a first-order HMM, using an enriched tagset that marks not only the positions of the named entities, but the tokens in front of and behind them. Two different strategies are employed for handling unknown tokens. The first (the `TokenShapeChunker`) replaces unknown or rare tokens with a morphologically-based classification. The second, newer module (the `CharLmHmmChunker`) estimates the probability of an observed word given a tag using language models based on character-level n-grams. The LingPipe developers suggest that the `TokenShapeChunker` typically outperforms the

[4] `http://www.alias-i.com/lingpipe/`

CharLmHmmChunker. However, the more sophisticated handling of unknown words by the CharLmHmmChunker suggests that it might be a good fit to the domain.

As well as examining the performance of LingPipe out of the box, we were also able to make some customisations. We have a custom tokeniser, containing several adaptations to chemical text. For example, our tokeniser will only remove brackets from the front and back of tokens, and only if that would not cause the brackets within the token to become unbalanced. For example, no brackets would be removed from '(*R*)-acetoin'. Likewise, it will only tokenise on a hyphen if the hyphen is surrounded by two lower-case letters on either side (and if the letters to the left are not common prehyphen components of chemical names), or if the string to the right has been seen in the training data to be hyphenated with a chemical name (e.g. 'derived' in 'benzene-derived'). By contrast, the default LingPipe tokeniser is much more aggressive, and will tokenise on hyphens and brackets wherever they occur.

The CharLmHmmChunker's language models can also be fed dictionaries as additional training data—we have experimented with using a list of chemical names derived from ChEBI (de Matos et al., 2006), and a list of chemical elements. We have also made an extension to LingPipe's token classifier, which adds classification based on chemically-relevant suffixes (e.g. -yl, -ate, -ic, -ase, -lysis), and membership in the aforementioned chemical lists, or in a standard English dictionary.

We analysed the performance of the different LingPipe configurations by 3-fold cross-validation, using the 42-paper corpus described in the previous section. In each fold, 28 whole papers were used as training data, holding out the other 14 as test data. The results are shown in Table 5.

From Table 5, we can see that the character n-gram language models offer clear advantages over the older techniques, especially when coupled to a custom tokeniser (which gives a boost to F of over 7%), and trained with additional chemical names. The usefulness of character-based n-grams has also been demonstrated elsewhere (Wilbur et al., 1999; Vasserman, 2004; Townsend et al., 2005). Their use here in an HMM is particularly apt, as it allows the token-internal features in the language model to be combined with the token context.

Configuration	P	R	F
TokenShape	67.0%	52.9%	59.1%
$+c$	71.2%	62.3%	66.5%
$+t$	67.4%	52.5%	59.0%
$+c+t$	73.3%	62.5%	67.4%
CharLm	62.7%	63.4%	63.1%
$+l$	59.8%	68.8%	64.0%
$+t$	71.1%	70.0%	70.5%
$+l+t$	75.3%	73.5%	74.4%

Table 5: LingPipe performance using different configurations. c = custom token classifier, l = chemical name lists, t = custom tokeniser

The impact of custom tokenisation upon the older TokenShapeChunker is less dramatic. It is possible that tokens that contain hyphens, brackets and other special characters are more likely to be unknown or rare tokens—the TokenShapeChunker has previously been reported to make most of its mistakes on these (Vlachos and Gasperin, 2006), so tokenising them is likely to make less of an impact. It is also possible that chemical names are more distinctive as a string of subtokens rather than as one large token—this may offset the loss in accuracy from getting the start and end positions wrong. The CharLmHmmChunker already has a mechanism for spotting distinctive substrings such as 'N,N'-' and '-3-', and so the case for having long, well-formed tokens becomes much less equivocal.

It is also notable that improvements in tokenisation are synergistic with other improvements—the advantage of using the CharLmHmmChunker is much more apparent when the custom tokeniser is used, as is the advantage of using word lists as additional training data. It is notable that for the unmodified TokenShapeChunker, using the custom tokeniser actually harms performance.

6 Conclusion

We have produced annotation guidelines that enable the annotation of chemicals and related entities in scientific texts in a highly consistent manner. We have annotated a corpus using these guidelines, an analysis of which, and the results of using an offthe-shelf NER toolkit, show that finding good approaches to tokenisation and the handling of unknown words is critical in the recognition of these entities. The corpus and guidelines are available by contacting the first author.

7 Acknowledgements

We thank Ann Copestake and Peter Murray-Rust for supervision, Andreas Vlachos and Advaith Siddharthan for valuable discussions, and David Jessop for annotation. We thank the RSC for providing the papers, and the UK eScience Programme and EPSRC (EP/C010035/1) for funding.

References

Peter T. Corbett and Peter Murray-Rust. 2006. High-Throughput Identification of Chemistry in Life Science Texts. *CompLife*, LNBI 4216:107–118.

P. de Matos, M. Ennis, M. Darsow, M. Guedj, K. Degtyarenko and R. Apweiler. 2006. ChEBI —Chemical Entities of Biological Interest. *Nucleic Acids Res*, Database Summary Paper 646.

George Demetriou and Rob Gaizauskas. 2003. Corpus resources for development and evaluation of a biological text mining system. *Proceedings of the Third Meeting of the Special Interest Group on Text Mining*, Brisbane, Australia, July.

Shipra Dingare, Malvina Nissim, Jenny Finkel, Christopher Manning and Claire Grover. 2005. A system for identifying named entities in biomedical text: how results from two evaluations reflect on both the system and the evaluations. *Comparative and Functional Genomics*, 6(1-2),77-85.

Nick Kemp and Michael Lynch. 1998. Extraction of Information from the Text of Chemical Patents. 1. Identification of Specific Chemical Names. *J. Chem. Inf. Comput. Sci.*, 38:544-551.

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.

Jin-Dong Kim, Tomoko Ohta, Yoshimasa Tsuruoka, Yuka Tateisi and Nigel Collier. 2004. Introduction to the Bio-Entity Recognition Task at JNLPBA. *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, 70-75.

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.

Alexander A. Morgan, Lynette Hirschman, Marc Colosimo, Alexander S. Yeh and Jeff B. Colombe. 2004. Gene name identification and normalization using a model organism database. *Journal of Biomedical Informatics*, 37(6):396-410.

Meenakshi Narayanaswamy, K. E. Ravikumar and K. Vijay-Shanker. 2003. A Biological Named Entity Recogniser. *Pac. Symp. Biocomput.*, 427-438.

Irena Spasic and Sophia Ananiadou. 2004. Using Automatically Learnt Verb Selectional Preferences for Classification of Biomedical Terms. *Journal of Biomedical Informatics*, 37(6):483-497.

Bingjun Sun, Qingzhao Tan, Prasenjit Mitra and C. Lee Giles. 2007. Extraction and Search of Chemical Formulae in Text Documents on the Web. *The 16th International World Wide Web Conference (WWW'07)*, 251-259.

Manabu Torii and K. Vijay-Shanker. 2002. Using Unlabeled MEDLINE Abstracts for Biological Named Entity Classification. *Genome Informatics*, 13:567-568.

Manabu Torii, Sachin Kamboj and K. Vijay-Shanker. 2004. Using name-internal and contextual features to classify biological terms. *Journal of Biomedical Informatics*, 37:498-511.

Joe A. Townsend, Ann A. Copestake, Peter Murray-Rust, Simone H. Teufel and Christopher A. Waudby. 2005. Language Technology for Processing Chemistry Publications. *Proceedings of the fourth UK e-Science All Hands Meeting*, 247-253.

Alexander Vasserman. 2004. Identifying Chemical Names in Biomedical Text: An Investigation of the Substring Co-occurence Based Approaches. *Proceedings of the Student Research Workshop at HLT-NAACL*. 7-12.

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

W. John Wilbur, George F. Hazard, Jr., Guy Divita, James G. Mork, Alan R. Aronson and Allen C. Browne. 1999. Analysis of Biomedical Text for Chemical Names: A Comparison of Three Methods. *Proc. AMIA Symp.* 176-180.

Alexander Yeh, Alexander Morgan, Marc Colosimo and Lynette Hirschman. 2005. BioCreAtIvE Task IA: gene mention finding evaluation. *BMC Bioinformatics* 6(Suppl I):S2.