Language Resources and Chemical Informatics

C.J. Rupp[1], Ann Copestake[1], Peter Corbett[2], Peter Murray-Rust[2], Advaith Siddharthan[1], Simone Teufel[1], Benjamin Waldron[1]

[1] Computer Laboratory, University of Cambridge
[2] Unilever Centre for Molecular Informatics, University of Cambridge

Abstract

Chemistry research papers are a primary source of information about chemistry, as in any scientific field. The presentation of the data is, predominantly, unstructured information, and so not immediately susceptible to processes developed within chemical informatics for carrying out chemistry research by information processing techniques. At one level, extracting the relevant information from research papers is a text mining task, requiring both extensive language resources and specialised knowledge of the subject domain. However, the papers also encode information about the way the research is conducted and the structure of the field itself. Applying language technology to research papers in chemistry can facilitate eScience on several different levels.

The SciBorg project sets out to provide an extensive, analysed corpus of published chemistry research. This relies on the cooperation of several journal publishers to provide papers in an appropriate form. The work is carried out as a collaboration involving the Computer Laboratory, Chemistry Department and eScience Centre at Cambridge University, and is funded under the UK eScience programme.

1. Language Resources for Scientific Text

We treat the text of the chemistry research papers as natural language populated to varying degrees by unfamiliar terms and notations, e.g.:

> Dialkyl 9-chloro-3*H*-pyrrolo[1,2-*a*]indole-2,3-dicarboxylates are obtained in excellent yields from the 1 : 1 : 1 addition reaction between triphenylphosphine, dialkyl acetylenedicarboxylates and 3-chloroindole-2-carbaldehyde; dimethyl 9-chloro-3*H*-pyrrolo[1,2-*a*]indole-2,3-dicarboxylate is converted to dimethyl 9-oxo-9*H*-pyrrolo[1,2-*a*]indole-2,3-dicarboxylate.

We do not address the problem of parsing chemistry as such but rather augment existing parsing tools, extending their coverage to chemistry as an example of scientific text.
We, therefore, require domain-independent analysis tools for English and specialised components for recognising and analysing chemical terms and notation. We adopt a multi-engine approach to analysing such a large corpus, to optimise both the quality and the coverage of the analyses. Multiple parsers are only complementary if they implement contrasting techniques and produce compatible results.
We make use of two parsing systems:

PET (Callmeier, 2002) is an efficient parsing system for HPSG grammars which is used with the ERG, English Resource Grammar (Flickinger, 2002). This is based on a detailed grammar and lexicon and provides analyses in the underspecified semantic formalism, RMRS. While the grammar and lexicon are handcoded, the disambiguation model is trained on a body of preferred analyses.

RASP (Briscoe et al., 2006) is a shallower analysis component based on a statistical CFG. It does not require a predefined lexicon, relying on POS tagging. RASP has been trained on a balanced corpus of English. Parsing results from the RASP system can take the form of trees or grammatical dependency relations. We convert RASP syntax trees to RMRS representations.

We, therefore, have a compatible parsing results and two existing parsing systems providing language resources for English, but we need to augment these with specialised analysis of the chemistry terms. The integration of results is carried out in an architecture that requires three key representations.

1.1. SciXML: XML Mark Up for Scientific Text

Fortunately, we have access to the XML markup of the chemistry papers provided by three major journal publishers. Each publisher's markup schema performs essentially the same function but includes individual customisations. We find it expedient to transform the markup to a common schema designed to represent the logical structure of scientific papers, SciXML (Rupp et al., 2006). With all the papers under a common schema, we know which SciXML elements contain text to be analysed. The transformation to SciXML can be performed by XSLT scripts and offers the possibility of extension of our tools to related domains. To this end, we have also defined an XSLT mapping from the NLM DTD, used for archiving PubMed papers, to SciXML.

```
<annot type='rasp_token' id='t2540'
from='18190' to='18195' deps='s87'
source='v2778' target='v2779' value='fatty'/>

<annot type='pos' id='p2327' from='18190'
to='18201' deps='ro69' value='NN2'/>

<annot type='oscar' id='ro69' from='18190'
to='18201' source='v2778' target='v2780' value='CM'/>
```

Figure 1: SAF annotations for token, type and NER information in XML format.

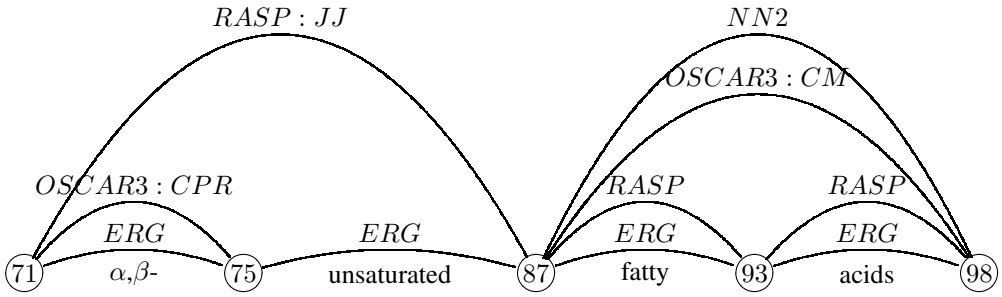

Figure 2: A chart for the tokenisation and tagging of: "α,β-unsaturated fatty acids"

1.2. SAF: Standoff Annotation Formalism

The combination of results from different analysers is founded on the ability to record results in a cumulative manner. In practice, we use SAF (Waldron and Copestake, 2006) standoff annotations which form a lattice of partial analyses at each level. We take the SciXML markup to be primary, indexing on that representation by character position. The efficiency of access to the lattice during processing is enhanced by encoding the annotations in an SQL database. For the more general representation and export of the information in the SAF annotations we use an XML format shown in Figure 1.

Figure 2 shows a partial SAF lattice containing information from various components in the SciBorg architecture, presented in Section 2.).

1.3. RMRS: Robust Underspecified Semantics

The robustness of Robust Minimal Recursion Semantics (Copestake, 2003) lies in the ability to represent varying degrees of resolution in semantic analyses. This allows us to work with the information available rather than generating redundant distinctions that cannot be substantiated. Put simply, "deep" (PET/ERG) and "shallow" (RASP), parsing results can be represented in the same form. In fact, even a sequence of lexical POS tags can be represented in an RMRS form.

1.4. OSCAR3: The recognition of Chemical Terms

Named entity recognition is a key problem in this system. As well as being of great interest to end users, named entities represent a source of words that may present difficulties to domain-independent parsing components. To assist with this, we have developed an extensive set of manual annotation guidelines, covering five classes of named entity. (Corbett et al., 2007) We have demonstrated that these guidelines can be applied to a range of chemistry papers with high inter-annotator agreement (F=0.93). We have studied the automated recognition of these entities using HMM-based systems. Significant gains were made by customisation of the tokenisation subsystems, the use of gazetteers of chemical names and the the use of character-level n-grams, allowing an F score for named entity recognition of 0.74. Current research suggests further improvements can be made via the use of Maximum Entropy Markov Models. The next step after detecting named entities is to assign semantics to them. In chemistry, this is complicated by a form of regular polysemy in which a chemical name can stand in for a specific compound, a class of compounds or a part of a compound. (Corbett et al., 2008) We have produced a set of manual annotation guidelines to address these ambiguities and have demonstrated acceptable inter-annotator agreement (86.0% accuracy, kappa = 0.784). A simple machine-learning system with a small feature set can make these distinctions with accuracy and kappa of 67.4% and 0.470.

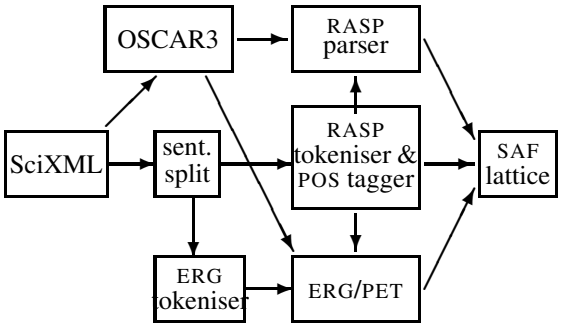

Figure 3: Parsing architecture of the SciBorg system

We can also assign chemical structures to named entities, partly by reference to databases and partly by machine interpretation of systematic chemical names (Corbett and Murray-Rust, 2006). This allows for new Information Retrieval techniques where keyword searches can be combined with searches for compounds containing particular structural motifs.

The utility of chemical named entity recognition has been demonstrated through our collaboration with the Royal Society of Chemistry, where our named-entity software OSCAR3 (Corbett and Murray-Rust, 2006) has been integrated into scientific publishing work-flows, to produce semantically-enriched chemistry papers in the award-winning Project Prospect system. (Batchelor and Corbett, 2007)

2. An Architecture for Integrating Results

To augment the two parsing systems with information provided by OSCAR3 about the form and function of chemical terms, we have factored them into their component modules. We do not alter the functionality of these modules, but we construct an integrated architecture which combines OSCAR3 results in the messages passed between the internal interfaces (see Figure 3). This architecture relies on the lattice of SAF annotations as a common representation of intermediate results. It also permits a closer integration of the two parser enhancing overall coverage and robustness.

The RASP system components provide the backbone of the architecture, as they instantiate each of the functions required for the complete parsing architecture: sentence splitter, tokeniser, POS tagger and syntactic parser. PET uses a distinct tokeniser and a core parsing module. OSCAR3 can be seen as providing a tokenisation, in recognising chemical terms, and a form of tagging in classifying their function.

As an example, the (partial) SAF lattice, represented in Figure 2, contains edges from the RASP tokeniser, the ERG tokeniser, OSCAR3 and the RASP POS tagger. The words of the input string appear between the corresponding lattice nodes. The OSCAR3 classification of the chemical terms also appears on those edge labels. Where the RASP token edge and POS edge coincide a single edge bears both labels. The second isolated POS edge arises from the selection of edges to submit to the POS tagger.

The RASP parser requires a single sequence of tagged tokens, so incorporating OSCAR3 terms amounts to selecting a path through the SAF lattice, based on confidence values from OSCAR3 and the RASP tagger. For PET we exploit an unknown word mechanism which can approximate lexical type information based on a tagging, or, by extension, an OSCAR3 classification. As PET allows multiple tokenisations in the parser input, also forming a lattice, we can submit both the OSCAR3 terms and RASP tags on this interface, to be used in the absence of a predefined lexical entry. This liberates the parser from its dependence on a static lexicon.

3. Discourse Analysis of Scientific Literature

In Sciborg, we are developing a discourse analysis of scientific papers that is based on the rhetorical role of citations, determination of scientific attribution for specific intellectual content and Argumentative Zoning (AZ).

3.1. Scientific Attribution

Scientific papers revolve around citations, and for many discourse level tasks one needs to know whose work is being talked about at any point in the discourse. For instance, in citation function classification (see Section 3.2.), the task is to find out if a citation is described as flawed or as useful. Consider:

> Most computational models of discourse are based primarily on an analysis of the intentions of the speakers **[Cohen and Perrault, 1979][Allen and Perrault, 1980][Grosz and Sidner, 1986]**$^{\text{WEAK}}$. The speaker will form intentions **based on** his goals and then act on these intentions, producing utterances. The hearer will then reconstruct a model of the speaker's intentions upon hearing the utterance. This approach has many strong points, but **does not provide a very satisfactory account** of the adherence to discourse conventions in dialogue.

In this example, the three citations are described as flawed (detectable by *"does not provide a very satisfactory account"*), but in order to find out, one must first realise that *this approach* refers to the three cited papers. A contrasting hypothesis could be that the citations are *used*; the cue phrase *"based on"* might make us think so (as in the context *"our work is based on"*). This, however, can be ruled out if we know that *the speaker* is not referring to some aspect of the current paper.

For other information access and retrieval purposes, the relevance of a citation within a paper can be crucial. One can estimate how important a citation is by simply counting how often it occurs in the paper. But as Kim and Webber (2006) argue, this ignores many expressions in text which refer to the cited author's work but which are not as easy to recognise as citations. In Siddharthan and Teufel (2007), we define the scientific attribution task in relation to standard anaphora resolution tasks and show that a range of linguistic expressions including definite descriptions and pronouns can be attributed to citations with Krippendorff's Alpha of 0.67 and percentage agreement greater than 85%. We further show that information about scientific attribution can be directly converted to features that boost the performance of our AZ classifier.

3.2. Citation Analysis

In Teufel et al. (2006), we describe a rhetorical scheme for annotating citation function. Our scheme has 12 categories, which broadly identify the citation as being supported or used, criticised, compared or contrasted, or just neutrally described:

- Agreement/usage/compatibility with other work (6 categories)
  - PBAS: basis or starting point
  - PUSE: usage of some aspect
  - PMODI: usage with modification
  - PMOT: motivating problem or choice of solution
  - PSIM: similarity of approach or goal
  - PSUP: support or compatibility
- WEAK: Explicit statement of weakness
- Contrast or comparison with other work (4 categories)
  - COCOGM: contrasts in methods or goals
  - COCOR0: neutral comparison of results
  - COCO-: superiority to cited work
  - COCOXY: contrasts between two different cited works
- NEUT: A neutral category.

Some examples follow to illustrate the categories:

> WEAK: *Most previous EIS studies of biological binding have used multilayer films or other complex structures, [cit17] [cit23] that have made it difficult to achieve a fundamental understanding of the electrical signal transduction process, particularly in the case of proteins.*

> COCO-: *The immediate impact of the dmphen ligand-support was that the reaction could be performed with 1 molpercent of Pd(ii), down from 10 molpercent in the absence of the ligand. [cit3d]*

PUSE: *The steady-state equilibrium binding constants and kinetic constants of corroles 3, 5 with duplex and quadruplex DNA were both measured under previously described experimental conditions. [cit16a] [cit16b]*

PSUP: *This result is consistent with electromagnetic theory [cit3] [cit8] where coupling and shifting of surface plasmons occur in aggregated particles relative to the case of an isolated particle.*

PMODI: *In our synthesis of 4-pyridinium corrole 3, we obtained this compound in two steps by a modification of the reported method. [cit12]*

Inter-annotator agreement was Kappa=.72 (12 classes;548 instances; 3 annotators)[1] and for automatic classification using features based on linguistic features, cue phrases and context, we got accuracy of 77% and Kappa=.57. Current work includes the incorporation of attribution information into our automatic rhetorical citation function analysis, and the adaptation of the entire discourse module to the chemistry domain.

3.3. AZ: Argumentative Zoning

Another use of discourse analysis for searching concerns the rhetorical status of entire sentences or larger segments. For instance, if we can identify that a particular sentence in a paper describes a *conclusion* rather than part of the *methodology*, this can be used for detection of contradictions between the main findings of papers. Lisacek et al. (2005) apply a similar thought to genetics papers.

AZ (Teufel, 2000; Teufel and Moens, 2002) is a theory describing high-level argumentation in scientific articles and how it relates to descriptions of the authors' own and other people's work.

An automatic recogniser for AZ exists, which is based on a set of 15 recognisable features and a machine learning component. Automatic annotation in Teufel and Moens (2002) achieved an agreement of 78% (Kappa=.45) with human gold standard annotation. AZ was originally devised for articles in the computational linguistics domain. However, we cannot simply assume that the AZ-recogniser for computational linguistics articles will perform well on chemistry articles.

In SciBorg, we port Argumentative Zoning on the basis of redefined categories. We have subdivided the OWN category (description of novel knowledge claim) into methods, conclusions and (objectively measurable) results. One other new search task we identified concerns the detection of failed problem solving activities, as in

OWN_FAIL: *Unfortunately, the observed low yields of the crystalline samples have prevented the use of NMR measurments.*

The automatic detection of such segments supports searches for synthesis methods which do not work, a common information need for synthetic chemists. Some deeper discourse analysis is necessary here, because not only do we need to identify that a failed problem solving activity took place, we also need to find out that it is not associated with other researchers, but with the paper authors themselves.

We have also introduced more detail by subdividing the BASIS and CONTRAST categories. The old BASIS category has been replaced by two separate categories for (a) usage of others' tools, products or methodology and (b) support for others' findings or theories:

USE: *The complexes were synthesised following the procedure previously reported (citation).*

SUPPORT: *This value is consistent with several literature reports for the first protonation constant of catechol (citations) included in Table 2.*

The old CONTRAST is now separate categories for (a) statements of weakness in other peoples's work, and gaps in the literature (b) statements contradicting the theory or findings of others and (c) neutral statements of comparsions and contrasts of current work to others':

GAP: *Benzotelluretes have to the best of our knowledge never been synthesized.*

ANTI_SUPPORT: *The combined results above suggest that leaching tests (citations) that have used DDW or aqueous solutions containing only inorganic anions (e.g., SPLP) likely have underestimated actual copper leaching from brake wear debris during rainfa ll events and/or in storm water runoff.*

COMPARISON: *In contrast, the measurements that we report here were obtained at the open circuit potential, where the net current flow is zero.*

Manual annotation with the redefined categories is currently underway for articles from several sub-areas of chemistry in our corpus.

4. Conclusion

We have described an ongoing research project that relies on combining generally available language resources with specialised knowledge of both chemistry and the nature of scientific literature. This project should yield an analysed corpus of chemistry research which will provide a resource for chemical informatics. Since citations can be mapped across the corpus it may also provide a resource for further study of the structure of the field. The methodology should be applicable to providing eScience resources in other domains in science.

5. Acknowledgements

We are very grateful to the Royal Society of Chemistry, Nature Publishing Group and the International Union of Crystallography for supplying papers. This work was funded by EPSRC (EP/C010035/1) with additional support from Boeing.

[1] Following Carletta (1996), we measure agreement in Kappa, which follows the formula $K = \frac{P(A)-P(E)}{1-P(E)}$ where P(A) is observed, and P(E) expected agreement. Kappa ranges between -1 and 1. K=0 means agreement is only as expected by chance. Generally, Kappas of 0.8 are considered stable, and Kappas of .69 as marginally stable, according to the strictest scheme applied in the field.

6. References

Colin Batchelor and Peter Corbett. 2007. Semantic enrichment of journal articles using chemical named entity recognition. In *Proceedings of the ACL*, pages 45–48, Prague, Czech Republic. Demo and Poster Sessions.

E. Briscoe, J. Carroll, and R. Watson. 2006. The second release of the rasp system. In *Proceedings of the COLING/ACL-06*, Sydney, Australia. Interactive Presentation Sessions.

Ulrich Callmeier. 2002. Pre-processing and encoding techniques in PET. In Stephan Oepen, Daniel Flickinger, Jun'ichi Tsujii, and Hans Uszkoreit, editors, *Collaborative Language Engineering: a case study in efficient gra mmar-based processing*. CSLI Publications, Stanford.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Ann Copestake. 2003. Report on the design of rmrs. DeepThought project deliverable.

Peter Corbett and Peter Murray-Rust. 2006. High-throughput identification of chemistry in life science texts. In *Proceedings of Computational Life Sciences*, pages 107–118, Cambridge, UK.

Peter Corbett, Colin Batchelor, and Simone Teufel. 2007. Annotation of chemical named entities. In *BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 57–64, Prague, Czech Republic.

Peter Corbett, Colin Batchelor, and Ann Copestake. 2008. Pyridines, pyridine and pyridine rings: disambiguating chemical named entities. In *Building and evaluating resources for biomedical text mining*, Marrakech, Morocco. Workshop at LREC.

Dan Flickinger. 2002. On building a more efficient grammar by exploiting types. In Stephan Oepen, Daniel Flickinger, Jun'ichi Tsujii, and Hans Uszkoreit, editors, *Collaborative Language Engineering: a case study in efficient gra mmar-based processing*, pages 1–17. CSLI Publications, Stanford.

Y. Kim and B. Webber. 2006. Automatic reference resolution in astronomy articles. In *In Proc. of 20th International CODATA Conference*, Beijing, China.

Frederique Lisacek, Christine Chichester, Aaron Kaplan, and Agnes Sandor. 2005. Discovering paradigm shift patterns in biomedical abstracts: Application to neurodegenerative diseases. In *Proc. of the SMBM*, European Bioinformatics Institute, Hinxton, UK.

CJ Rupp, Ann Copestake, Simone Teufel, and Ben Waldron. 2006. Flexible interfaces in the application of language technology to an escience corpus. In *Proceedings of the 4th UK E-Science All Hands Meeting*, Nottingham, UK.

Advaith Siddharthan and Simone Teufel. 2007. Whose idea was this and why does it matter? attributing scientific work to citations. In *Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT 2007)*, Rochester, New York.

Simone Teufel and Marc Moens. 2002. Summarising scientific articles — experiments with relevance and rhetorical status. *Computational Linguistics*, 28(4):409–446.

Simone Teufel, Advaith Siddharthan, and Dan Tidhar. 2006. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP'06)*, Sydney, Australia.

Simone Teufel. 2000. *Argumentative Zoning: Information Extraction from Scientific Text*. Ph.D. thesis, School of Cognitive Science, University of Edinburgh, Edinburgh, UK.

Benjamin Waldron and Ann Copestake. 2006. A Standoff Annotation Interface between DELPH-IN Components. In *The fifth workshop on NLP and XML: Multi-dimensional Markup in Natural Language Processing (NLPXML-2006)*.