

Language Technology for Processing Chemistry Publications

Joe Townsend¹, Ann Copestake², Peter Murray-Rust¹, Simone Teufel², Chris Waudby^{1,3}

1. Unilever Centre for Molecular Sciences Informatics, Department of Chemistry, University of Cambridge, CB2 1EW

2. Computer Laboratory, University of Cambridge, CB3 0DF

3. Department of Chemistry, University of Cambridge, CB2 1EW

Abstract

Extraction of chemical information embedded in primary publications is essential for the development of eScience. We describe text mining tools and procedures which can identify, understand and extract such information on a high-throughput basis.

1. Introduction

In many scientific domains the data on which eScience depends are not published coherently but scattered through tens of thousands of independent journal articles.

The synthesis and properties of over 1 new million chemical compounds are published annually in an uncoordinated free-text form. This information is critical to many sciences and industry and is the basis of a \$500+ million chemical informatics business, based on human abstractors. Chemical compounds are identified by name, formula and pictorial depiction, then translated into a machine representation ("connection table", CT) and given a unique identifier. Details of recipes and procedures are also manually abstracted, though in less detail and consistency.

This is a major opportunity for high-throughput text-mining which can not only lower the cost but also increase the range of information extracted, especially for the details of chemical recipes. We report the enhancement of our toolset (OSCAR [Adams, S. E., J. M. Goodman, et al., 2004 and Townsend, J. A., S. E. Adams, et al., 2004]) and introduce the use of more generic parsing tools (RMRS, RASP [Briscoe and Carroll, 2002]).

2. Chemical articles

From the analysis of several hundred articles we find that chemical syntheses are reported in implicit conventional style, whose structure (apart from the bibliographic and citation information) regularly consists of:

Abstract. Understanding text and extracting text from abstracts is hard as there is little context. This will require deeper language processing techniques (cf. section 4). Chemical entity recognition may be possible.

Introduction. This may include previous work, rationale, and general methodology but has no required structure and is often small. The most accessible content is motivation (shown below) rather than precise information extraction:

In previous parts of this series we have reported on synthetic routes to pyrido- and quino-acridine systems ... Recently we have reported the synthesis of the intriguing pentacyclic acridinium salt 3... the design of inhibitors of this enzyme is of burgeoning interest to anticancer drug design teams.... Our original route to tetracyclic systems

involved ... this route is inefficient in practice,... we return to the problem of synthesising 5-substituted triazoles... [Ellis and Stevens, 2001]

As shown by the highlighted phrases the authors are clearly developing their own method incrementally, also, introduction sections often contain mentions of other researchers' work and can thus be relevant to citation processing (cf. section 4).

Methodology. This often uses formulaic language and may contain descriptions of substances, instruments, recipes and observations, and this typical paragraph is composed almost entirely of named entities and stock phrases:

NMR spectra were recorded on a Bruker ARX 250 spectrometer at room temperature. Chemical shifts are reported in δ units and referenced to the solvent as internal standard; coupling constants (J values) are in Hz. Melting points were determined on a Gallencamp melting point apparatus and are uncorrected. IR spectra were measured on a Mattson 2020 Galaxy Series FT-IR spectrometer, UV spectra on a Pharmacia Biotech Ultraspec 2000 UV/visible spectrometer and mass spectra on a Micromass Platform spectrometer. High resolution mass data were collected on a VG Autospec instrument. Differential scanning calorimetry was performed with a Perkin-Elmer Pyris 1 instrument as previously reported.²⁻⁴ Merck silica gel 60 (0.04-0.63 mm) was used for chromatography. [Ellis and Stevens, 2001]

Terms are precise, and often repeated verbatim from paper to paper so regular expressions can be used to populate lexicons which will then also give valuable current and historic analysis of practice.

Experimental. This is highly ritualised [Fig 1] (and could be better represented by structured data files). It consists of a stylised recipe followed by textual reports of data items including molecular properties and spectra. This information is of great value to the community and databases with this content are resold for drug and materials design. The journal guidelines are almost a regular grammar, and well suited to parsing by regular expression, but authors often introduce minor errors making it non-deterministic.

Results and discussion. This can describe the success and failure of reactions:

*Brief treatment of new triazoles **11b-g** in boiling diphenyl ether generally gave new 2-substituted **14** in acceptable (30 - 90%) yields. However, the pyrrolidinyl derivative **14c** and the **14g** could not be isolated. ... The precursor **16** could not be prepared from the chloromethyltriazole **8** and indoline without competitive nucleophilic substitution at the acridine 9-position. However, conversion of the chloroacetyltriphenyl-phosphorane ylide **7** to the indolinyl analogue **15** proceeded smoothly (73%);* [Ellis and Stevens, 2001]

Reports of failure (as well as success) are of great value to the systematisation of synthetic chemistry.

3. Chemical text-mining. Despite the value and tractability of machine analysis of chemical papers, there is almost no current published work. We identify several topics:

Classification of documents and their content, through named entity occurrence and co-occurrence. Where documents map to concepts (organisms, diseases, methodology) it may be possible to classify those concepts. This is not discussed further here.

Lexicons. Manual analysis of 20 papers from *Organic & Biomolecular Chemistry* indicated the need for chemistry-specific lexicons for the following concepts in recipes.

Actions	Apparatus	Conditions	Descriptions
Groups	Manners	Methods	Objects
Quantities	ProcedureReferences	States	SubstanceReferences
Substances	Times		

Apart from manual curation these can also be populated by shallow parsing. The resulting lexicons are of considerable interest as scientific objects for analysing chemical synthetic and analytic procedures.

Lexical identification of chemical entities. Chemical names (e.g. *tetrahydroginestrone*) can be recognised as such even if their interpretation is unclear. For the training set, a corpus of 7693 chemical names (C) was created using data from MSD [EMBL-EBI 2004] and 20895 English words (E) without chemical names was created from papers in *Nature*.

The original toolkit [Adams, S. E., J. M. Goodman, et al. 2004] used regular expression-based pattern matching with word starts and endings (*acet-*, *acr-*, *ada-*, *-ace*, *-al*, *-ane*, etc) with a lexicon of stop words (e.g. animal, arose). Although moderately successful, this approach was limited by the tremendous variety of chemical names in the literature.

We extended this through the naïve Bayes method applied to overlapping 4-Grams {aaaa, aaab, ..., zzzz}; thus *methanol* contains the feature set ($\{w_j\}$, all contained in the word W) { m,m , me,m , met,m , *meth*, *etha*, *than*, *hano*, *anol*, *no*\$, *ol*\$\$, *l*\$\$\$}. We let C represent the category of chemical names, and E that of English words (identical to the complement of C).

$$\frac{P(C|W)}{P(E|W)} = \gamma \cdot \prod_j \frac{P(w_j|C)}{P(w_j|E)}; \gamma = \frac{P(C)}{1 - P(C)}.$$

γ is a parameter which reflects the abundance of chemical names in text. We also use context to adjust the probability score, e.g. to distinguish *periodic acid* from *periodic table*.

log γ	General text		Experimental	
	P (%)	R (%)	P (%)	R (%)
-2	68.5	94.3	85.9	95.3
-5	72.1	92.5	93.8	95.3
-8	75.0	90.6	96.8	95.3
-11	81.0	88.7	98.4	95.3
-14	80.8	79.2	98.2	87.5

Table 1 - Effect of γ on recall and precision in general text and experimental sections calculated using 10 articles from *Organic & Biomolecular Chemistry*.

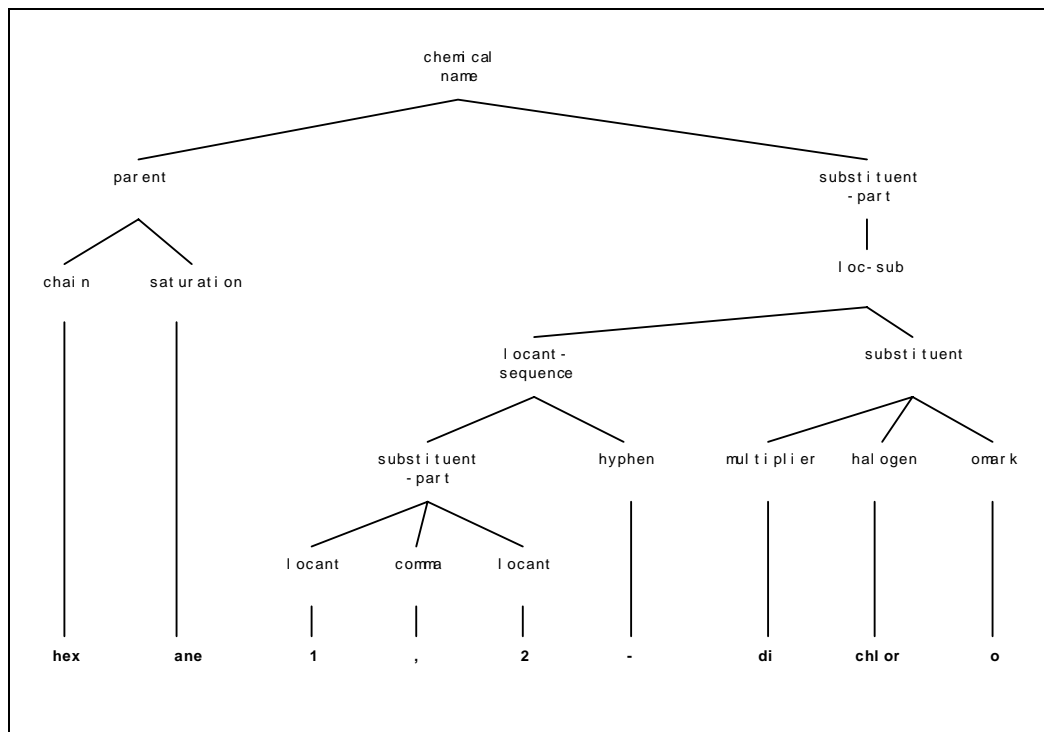
Interpretation and of chemical entities, since chemical names may be semantically rich it is often possible not only to recognise their form but also interpret their precise meaning. Thus the molecular structure (CT) of *butyl-chloride* is algorithmically deducible from IUPAC rules, while *tetrahydroginestrone*, while clearly a chemical, requires a lexicon for its CT.

Many chemical names are formed in a regular manner, such as **1, 2 - dichlorohexane**.

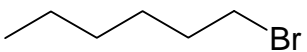
Lexical and syntactic analysis creates the following tokens.

locant 1; Comma; locant 2; Hyphen; multiplier di; halogen chlor; omark; chain hex; saturation
ane

The parse tree (below) leads to the molecular connection table.



However many names are not parsable in a deterministic manner, being "trivial" (e.g. *penicillanic acid*), partially regular (*2-chlorotestosterone*) or ambiguous (*2-chloroethylbenzene*). We tackle the continual creation of new chemical lexemes by adding names and structures to the lexicon.

Thus coupling  to *1-bromo-hexane* interprets "hexane" and its locants so that *1-chloro-hexane* could subsequently be interpreted.

However in some cases the ambiguities require probabilistic methods. *2-chloroethyl benzene* has insufficient locants and could mean *1-chloro-2-ethyl benzene* or *(2-chloro-ethyl) benzene*.

Disambiguation is problematic and requires context, although experimental information (e.g. spectra) could provide constraints.

Parsing recipes. This discourse is important, ritualised and is partially tractable through shallow parsing, local syntax and entity recognition (Townsend, J. A., S. E. Adams, et al., 2004 coloured).

n-Butyllithium 2.5 M in hexane (4.8 mL, 12 mmol) was added to a solution of 1,4-dibromo-2,5-dimethoxybenzene (4) (1.001 g, 30 mmol) in dry THF (20 mL) at -78 °C under a nitrogen atmosphere and stirred for 2 h at the same temperature. To this solution was added a further portion of dry THF (20 mL). To the reaction mixture was added 3.0 mL (39 mmol) of DMF and the solution was stirred for 60 min and hydrolysed with 10 ml 3 N hydrochloric acid. The reaction mixture was allowed to warm to room temperature. The yellow precipitate was filtered by suction. After drying *in vacuo*, yellow crystals of the dicarboxaldehyde (5) (0.4 g, 60.6%) were obtained. [Kuhnert, Rossignolo et al., 2003]

Figure 1

We represent it as *steps*, many of which represent chemical reactions, whose products are often identified (4, 5) but also anonymous ("yellow crystals"). Anaphora ("to this solution") can be problematic. Step recognition uses frequent tokens such as stops, 'and', 'then', giving

```
<stepList><step type="add"><reagent><substance>n-Butyllithium</substance> (<property type="amount">2.5 M</property>) <solvent>in <substance>hexane</substance> <property type="amount">(4.8 mL, 12 mmol)</property></solvent></reagent> <action>was added to</action> <reagent>a solution of <substance>1,4-dibromo-2,5-dimethoxybenzene (4)</substance> <property type="amount">(1.001 g, 30 mmol)</property> <solvent>in<state>dry</state> <substance>THF</substance> <property type="amount">(20 mL)</property></solvent></reagent> <conditions>at <quantity>-78°C</quantity></conditions> <unknown>under</unknown> <reagent>a <substance>nitrogen</substance></reagent> <unknown>atmosphere</unknown></step> and <step type="stir"><action>stirred for</action> <conditions><quantity>2 h</quantity></conditions> <conditions>at the same temperature</conditions></step>....
```

Information extraction. The analytical data and compound properties are very regular

Mp 193–195 °C; ν_{\max} (Nujol)/cm⁻¹ 1682 (C=O), 1482–1377 (C_{Ar}=C_{Ar}), 1215 (C–O), 877 (isolated Aryl-H); δ_{H} (270 MHz; CDCl₃) 10.5 (2H, s, CHO), 7.5 (2H, s, H-3 and H-6), 4.0 (3H, s, OCH₃); δ_{C} (270 MHz; CDCl₃) 189.6, 156.0, 129.4, 111.1, 56.3; *m/z* (EI) 194 (100%, M⁺), 166 (80%, M-CO); CHN (Found: C, 61.4; H, 5.20. C₁₀H₁₀O₄ requires: C, 61.8; H, 5.19%).

and have good recall and precision:

Data Type	Recall (%)	Precision	
		(%)	F (%)
HRMS	99.2	100.0	99.6
Melting Point	93.2	98.7	95.9
C-NMR	93.0	97.4	95.2
H-NMR	90.2	98.2	94.1
Mass Spec	87.9	100.0	93.7
Elemental Analysis	87.3	100.0	93.4

Infra Red	90.7	95.9	93.3
Name (OSCAR 1)	70.4	78.4	74.3

$$R = \frac{TP}{TP + FN} \quad P = \frac{TP}{TP + FP} \quad F = \frac{2PR}{P + R}$$

Table 2. *TP* = true positives; *FP* = false positives; *FN* = false negatives; *R* = recall; *P* = precision.

Extraction of this data is extremely valuable as spectral quantities can generate new scientific insights and confirm identities of compounds.

4. Future directions.

While the work described above demonstrates that we can successfully deal with some portions of Chemistry texts, it has clear limitations. We propose to address the more open-ended text processing tasks by using more general NLP technology. We will adapt and develop shallow and deep parsing technology in a project that is due to start in October 2005. ('Extracting the Science from Scientific Publications', EPSRC EP/C010035/1, Computer Science Challenges to Emerge e-Science: collaboration between the Computer Laboratory, Unilever Centre for Molecular Informatics, eScience Centre at Cambridge. The Royal Society of Chemistry, Nature Publishing Group and International Union of Crystallography are partners.)

Our approach is based on the idea of developing a natural-language oriented markup language which is compatible with Grid protocols, but which also enables the tight integration of partial information from a wide variety of language processing tools and has a sound logical basis compatible with Semantic Web standards. Robust Minimal Recursion Semantics (RMRS, [Copestake 2003]) is an application-independent representation which captures the information that comes from the syntax and morphology of natural language. (RMRS was originally developed on the EU-funded 'Deep Thought' project funded under the Thematic Programme User-friendly Information Society of the 5th Framework Programme of the European Community (Contract N° IST-2001-37836).) Crucially, RMRS is compatible with both shallow and deep language processing techniques, allowing us to investigate integrating a range of approaches to analysing texts. RMRS has been designed to be suitable for natural language representation and as such has to be very expressive while at the same time allowing for underspecification. Formally, RMRSs are partial descriptions which correspond to a set of logical forms in a higher-order base language. RMRSs can be linked to ontologies, so that the notion of underspecification of an RMRS reflects the hierarchical ontological relationship. RMRS is thus distinct from RDF/OWL, but there are interesting formal correspondences. For the applications in this project, such as various types of information extraction, RDF/OWL terms will be extracted from RMRSs.

RMRS marked-up text will be fully compatible with Chemical Markup Language (CML). We intend to integrate the domain-specific processing techniques described in the previous sections with domain-independent analysis of various levels using RMRS. For instance, chemical names can be regarded as a very specialised form of named entity. The domain-independent processors thus only need a general lexical entry pattern for chemical names which will enable them to analyse running text that incorporate them.

We believe that RMRS markup will allow us to extract richer and more varied information from texts than is possible with existing techniques. One example of this is scientific argumentation

structure [Teufel and Moens, 2002]. Knowledge of the overall discourse structure of the scientific text and interpretation of citation context can enhance human browsing and support more fine-grained searches in the literature. For instance, information about where the differences in two approaches lie and which researchers form a “school of thought” is one of the planned applications of this more open-ended text understanding task.

We plan to incrementally enhance the technology described above by incorporating more general NLP tools, all of which will produce RMRS. We aim to develop a toolkit that supports knowledge base acquisition, ontology construction and free-style browsing. This will aid chemistry researchers, both in mining the existing scientific record and in supporting authoring of e-publications with appropriate links and annotation. Longer term, we hope to use the same sort of blend of general purpose NLP technology and domain-specific techniques in other scientific disciplines.

5. References

- Adams, S. E., J. M. Goodman, et al. (2004). "Experimental data checker: better information for organic chemists." Org Biomol Chem **2**(21): 3067-70.
- Chen, S. F. and J. M. Goodman (1999). "An empirical study of smoothing techniques for language modelling." Computer Speech and Language **13**: 359-394.
- Copestake, A. 2003. Report on the design of RMRS. DeepThought project deliverable.
- EMBL-EBI (2004). Macromolecular Structure Database, www.ebi.ac.uk/msd.
- Ellis, M.J., Stevens, M.F.G J. Chem. Soc., Perkin Trans. 1, 2001, 3174–3179
- Goodman, J. T. (2001). A Bit of Progress in Language Modelling. Redmond, WA., Machine Learning and Applied Statistics Group, Microsoft Research.
- Jurafsky, M. and J. H. Martin (2000). Chapter 6. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. New Jersey, Prentice Hall.
- Kuhnert, N., G. M. Rossignolo, et al. (2003). "The synthesis of trianglimines: on the scope and limitations of the [3 + 3] cyclocondensation reaction between (1R,2R)-diaminocyclohexane and aromatic dicarboxaldehydes." Org Biomol Chem **1**(7): 1157-1170.
- Press, W. H., S. A. Teukolsky, et al. (1992). Numerical Recipes in FORTRAN 77, Cambridge University Press.
- Teufel, S. and Moens, M. (2002). Summarizing Scientific Articles - Experiments with Relevance and Rhetorical Status. In Computational Linguistics, 28 (4): 409-445.
- Townsend, J. A., S. E. Adams, et al. (2004). "Chemical documents: machine understanding and automated information extraction." Org Biomol Chem **2**(22): 3294-300.
- Vasserman, A. (2004). "Identifying chemical names in biomedical text: an investigation of the substring co-occurrence based approaches." Proceedings of the Student Research Workshop at HLT-NAACL.