Semantic enrichment of journal articles using chemical named entity recognition

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

Peter T. Corbett
Unilever Centre for Molecular Science Informatics
University Chemical Laboratory
Lensfield Road
Cambridge
UK CB2 1EW
ptc24@cam.ac.uk

Abstract

We describe the semantic enrichment of journal articles with chemical structures and biomedical ontology terms using Oscar, a program for chemical named entity recognition (NER). We describe how Oscar works and how it can been adapted for general NER. We discuss its implementation in a real publishing workflow and possible applications for enriched articles.

1 Introduction

The volume of chemical literature published has exploded over the past few years. The crossover between chemistry and molecular biology, disciplines which often study similar systems with contrasting techniques and describe their results in different languages, has also increased. Readers need to be able to navigate the literature more effectively, and also to understand unfamiliar terminology and its context. One relatively unexplored method for this is semantic enrichment. Substructure and similarity searching for chemical compounds is a particularly exciting prospect.

Enrichment of the bibliographic data in an article with hyperlinked citations is now commonplace. However, the actual scientific content has remained largely unenhanced, this falling to secondary services and experimental websites such as GoPubMed (Delfs *et al.*, 2005) or EBIMed (Rebholz-Schuhmann *et al.*, 2007). There are a few examples of semantic enrichment on small (a few dozen articles per year) journals such as *Nature Chemical Biology* being an example, but for a larger journal it is impractical to do this entirely by hand.

This paper concentrates on implementing semantic enrichment of journal articles as part of a publishing workflow, specifically chemical structures and biomedical terms. In the Motivation section, we introduce Oscar as a system for chemical NER and recognition of ontology terms. In the Implementation section we will discuss how Oscar works and how to set up ontologies for use with Oscar, specifically GO. In the Case study section we describe how the output of Oscar can be fed into a publishing workflow. Finally we discuss some outstanding ambiguity problems in chemical NER. We also compare the system to EBIMed (Rebholz-Schuhmann *et al.*, 2007) throughout.

2 Motivation

There are three routes for getting hold of chemical structures from chemical text—from chemical compound names, from author-supplied files containing connection tables, and from images. The preferred representation of chemical structures is as diagrams, often annotated with curly arrows to illustrate the mechanisms of chemical reactions. The structures in these diagrams are typically given numbers, which then appear in the text in bold face. However, because text-processing is more advanced in this regard than image-processing, we shall concentrate on NER, which is performed with a system called Oscar. A preliminary overview of the system was presented by Corbett and Murray-Rust (2006). Oscar is open source and can be downloaded from `http://oscar3-chem.sourceforge.net/`

As a first step in representing biomedical content, we identify Gene Ontology (GO) terms in full text.[1] (The Gene Ontology Consortium, 2000) We have chosen a relatively simple starting point in order to gain experience in implementing useful semantic markup in a publishing workflow without a substantial word-sense disambiguation effort. GO terms are largely compositional (Mungall, 2004), hence incomplete matches will still be useful, and that there is generally a low level of semantic ambiguity. For example, there are only 133 single-word GO terms, which significantly reduces the chance of polysemy for the 20000 or so others. In contrast, gene and protein

[1]We also use other OBO ontologies, specifically those for nucleic acid sequences (SO) and cell type (CL).

Proceedings of the ACL 2007 Demo and Poster Sessions, pages 45–48,
Prague, June 2007.

`(.*) activity$`	$\rightarrow$	`(\1)`
`(.*) formation$`	$\rightarrow$	$\emptyset$
`(.*) synthesis$`	$\rightarrow$	$\emptyset$
`ribonuclease`	$\rightarrow$	`RNAse`
	$\rightarrow$	`ribonuclease`
`^alpha-` (*etc.*)	$\rightarrow$	α- (*etc.*)
	$\rightarrow$	`alpha-` (*etc.*)
pluralize nouns		
stopwords	$\rightarrow$	$\emptyset$

Table 1: Example rules from 'Lucinda', used for generating recogniser input from OBO files

names are generally short, non-compositional and often polysemous with ordinary English words such as Cat or Rat.

3 Implementation

Oscar is intended to be a component in larger workflows, such as the Sciborg system (Copestake *et al.*, 2006). It is a shallow named-entity recogniser and does not perform deeper parsing. Hence there is no analysis of the text above the level of the term, with the exception of acronym matching, which is dealt with below, and some treatment of the boldface chemical compound numbers where they appear in section headings. It is optimized for chemical NER, but can be extended to handle general term recognition. The EBIMed system, in contrast, is a pipeline, and lemmatizes words as part of a larger workflow.

To identify plurals and other variants of non-chemical NEs we have a ruleset, nicknamed Lucinda, outlined in Table 1, for generating the input for the recogniser from external data. We use the plain-text OBO 1.2 format, which is the definitive format for the dissemination of the OBO ontologies.

We strive to keep this ruleset as small as possible, with the exception of determining plurals and a few other regular variants. The reason for keeping plurals outside the ontology is that plurals in ordinary text and in ontologies can have quite different meanings.

There is also a short stopword list applied at this stage, which is different from Oscar's internal stopword handling, described below.

3.1 Named entity recognition and resolution

Oscar has a recogniser to identify chemical names and ontology terms, and a resolver which matches NEs to ontology IDs or chemical structures. The recogniser classifies NEs according to the scheme in Corbett *et al.* (2007). The classes which are relevant here are `CM`, which identifies a chemical compound, either because it appears in Oscar's chemical dictionary, which also contains structures and InChIs,[2] or according to Oscar's n-gram model, regular expressions and other heuristics and `ASE`, a single word ending in "-ase" or "-ases" and representing an enzyme type. We add the class `ONT` to these, to cover terms found in ontologies that do not belong in the other classes, and `STOP`, which is the class of stopwords.

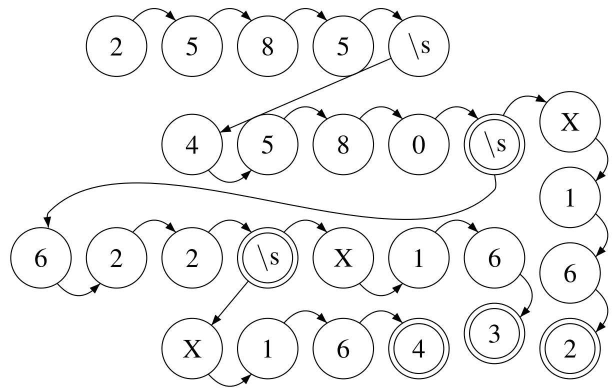

Figure 1: Cartoon of part of the recogniser. The mapping between this automaton and example GO terms is given in Table 2.

GO term	Regex pair
bud neck	`2585\s4580\s`
	`2585\s4580\sX162`
bud neck polarisome	`2585\s4580\s622\s`
	`2585\s4580\s622\sX163`
polarisome	`622\s`
	`622\sX164`

Table 2: Mapping in Fig. 1. The regexes are purely illustrative. IDs 162, 163 and 164 map on to GO:0005935, GO:0031560 and GO:0000133 respectively.

We sketch the recogniser in Fig. 1. To build the recogniser: Each term in the input data is tokenized and the tokens converted into a sequence of digits followed by a space. These new tokens are concatenated and converted into a pair of regular expressions. One of these expressions has X followed by a term ID appended to it. These regex–regex pairs are converted into finite automata, the union of which is determinized. The resulting DFA is examined for accept states. For each accept state for which a transition to X is also present, the sequences of digits after the X is used to build a mapping of accept states to ontology IDs (Table 2).

To apply the recogniser: The input text is tokenized, and for each token a set of representations is calculated which map to sequences of digits as above. We then make an empty set of DFA instances (a pointer to the DFA,

[2] An InChI is a canonical identifier for a chemical compound. `http://www.iupac.org/inchi/`

which state it's in and which tokens it has matched so far), and for each token, add a new DFA instance for each DFA, and for each representation of the token, clone the DFA instance. If it does not accept the digit-sequence representation of the token, throw it away. If it is in an accept state, note which tokens it has matched, and if the accept state maps to an ontology ID (ontID), we have an NE which can be annotated with the ontID.

Take all of the potential NEs. For all NEs that have the same sequence of tokens, share all of the ontIDs. Assign its class according to a priority list where `STOP` comes first and `CM` precedes `ASE` and `ONT`. For the system in Fig. 1, the phrase "bud neck polarisome" matches three IDs. We choose the longest–leftmost sequence. If the resolver generates an InChI for an NE, we look up this InChI in ChEBI (de Matos *et al.*, 2006), a biochemical ontology, and take the ontology ID. This has the effect of aligning ChEBI with other databases and systematic nomenclature.

3.2 Gene Ontology

In working out how to mine the literature for GO terms, we have taken our lead from the domain experts, the GO editors and the curators of the Gene Ontology Annotation (GOA) database.

The Functional Curation task in the first BioCreative exercise (Blaschke *et al.*, 2005) is the closest we have found to a systematic evaluation of GO term identification. The brief was to assign GO annotations to human proteins and recover supporting text. The GOA curators evaluated the results (Camon *et al.*, 2005) and list some common mistakes in the methods used to identify GO terms. These include annotating to obsolete terms, predicting GO terms on too tenuous a link with the original text, for example in one case the phrase "pH value" was annotated to "pH domain binding" (GO:0042731), difficulties with word order, and choosing too much supporting text, for example an entire first paragraph of text.

So at the suggestion of the GO editors, Oscar works on exact matches to term names (as preprocessed above) and their exact (within the OBO syntax) synonyms.

The most relevant GO terms to chemistry concern enzymes, which are proteins that catalyse chemical processes. Typically their names are multiword expressions ending in "-ase". The enzyme A B Xase will often be represented by GO terms "A B Xase activity", a description of what the enzyme does, and "A B Xase complex", a cellular component which consists of two or more protein subunits. In general the bare phrase "A B Xase" will refer to the activity, so the ruleset in Table 1 deletes the word "activity" from the GO term.

We shall briefly compare our method with the algorithms in EBIMed and GoPubMed. The EBIMed algorithm for GO term identification is very similar to ours, except for the point about lemmatization listed above, and its explicit variation of character case, which is handled in Oscar by its case normalization algorithm. In contrast, the algorithm in GoPubMed works by matching short 'seed' terms and then expanding them. This copes with cases such as "protein threonine/tyrosine kinase activity" (GO:0030296) where the full term is unlikely to be found in ordinary text; the words "protein" and "activity" are generally omitted. However, the approach in (Delfs *et al.*, 2005) cannot be applied blindly; the authors claim for example that "biosynthesis" can be ignored without compromising the reader's understanding. In chemistry journal articles most mentions of a chemical compound will not refer to how it is formed in nature; they will refer to the compound itself, its analogues or other processes. In fact, our ruleset in Table 1 explicitly disallows GO term synonyms ending in " synthesis" or " formation" since they do not necessarily represent biological processes. It is also not clear from Delfs *et al.* (2005) how robust the algorithm is to the sort of errors identified by Camon *et al.* (2005).

4 Case study

The problem is to take a journal article, apply meaningful and useful annotations, connect them to stable resources, allow technical editors to check and add further annotations, and disseminate the article in enriched form.

Most chemical publishers use XML as a stable format for maintaining their documents for at least some stages of the publication process. The Sciborg project (Copestake *et al.*, 2006) and the Royal Society of Chemistry (RSC) use SciXML (Rupp *et al.*, 2006) and RSC XML respectively. For the overall Sciborg workflow, standoff annotation is used to store the different sets of annotations. For the purposes of this paper, however, we make use of the inline output of Oscar, which is SciXML with `<ne>` elements for the annotations.

Not all of the RSC XML need be mined for NEs; much of it is bibliographic markup which can confuse parsers. Only the useful parts are converted into SciXML and passed to Oscar, where they are annotated. These SciXML annotations are then pasted back into the RSC XML, where they can be checked by technical editors. In running text, NEs are annotated with an ID local to the XML file, which refers to `<compound>` and `<annotation>` elements in a block at the end, which contain chemical structure information and ontology IDs. This is a lightweight compromise between pure standoff and pure inline annotation.

We find useful annotations by aggressive thresholding. The only classes which survive are `ONT`s, and those `CM`s which have a chemical structure found by the resolver. This enables the chemical NER part of Oscar to be tuned for high recall even as part of a publishing workflow. Only CMs which correspond to an unambiguous molecule or molecular ion are treated as a chemical compound; everything else is referred to an appropriate ontology. We use the InChI as a stable representation for chemical structure, and the curated OBO ontologies for biomedical terms.

The role of technical editors is to remove faulty annotations, add new compounds to the chemical dictionary, based on chemical structures supplied by authors, suggest new GO terms to the ontology curators, and extend the stopword lists of both Oscar and Lucinda as appropriate. At present (May 2007), this happens after publication of articles on the web, but is intended to become part of the routine editing process in the course of 2007.

This enriched XML can then be converted into HTML and RSS by means of XSL stylesheets and database lookups, as in the RSC's Project Prospect.[3] The immediate benefits of this work are increased readability of articles for readers and extensive cross-linking with other articles that have been enhanced in the same way. Future developments could easily involve structure-based searching, ontology-based search of journal articles, and finding correlations between biological processes and small molecule structures.

5 Ambiguity in chemical NER

One important omission is disambiguating the exact referent of a chemical name, which is not always clear without context. For example "the pyridine **6**", is a class description, but the phrase "the pyridine molecule" refers to a particular compound. ChEBI, which contains an ontology of molecular structure, uses plurals to indicate chemical classes, for example "benzenes", which is often, but not always, what "benzenes" means in text. Currently Oscar does not distinguish between singular and plural.

Amino acids and saccharides are particularly troublesome on account of homochirality. Unless otherwise specified, "histidine" and "ribose" specify the molecules with the chirality found in nature, or to be precise, L-histidine and D-ribose respectively. What is even worse is that "histidine" seldom refers to the independent molecule; it usually means the histidine residue, part of a larger entity.

6 Acknowledgements

We thank Dietrich Rebholz-Schuhmann for useful discussions. CRB thanks Jane Lomax, Jen Clark, Amelia Ireland and Midori Harris for extensive cooperation and help, and Richard Kidd, Neil Hunter and Jeff White at the RSC. PTC thanks Ann Copestake and Peter Murray-Rust for supervision. This work was funded by EPSRC (EP/C010035/1).

[3] http://www.projectprospect.org/

References

Christian Blaschke, Eduardo Andres Leon, Martin Krallinger and Alfonso Valencia. 2005. Evaluation of BioCreAtIvE assessment of task 2 *BMC Bioinformatics* 6(Suppl 1):S16

Evelyn B. Camon, Daniel G. Barrell, Emily C. Dimmer, Vivian Lee, Michele Magrane, John Maslen, David Binns and Rolf Apweiler. 2005. An evaluation of GO annotation retrieval for BioCreAtIvE and GOA *BMC Bioinformatics* 6(Suppl 1):S17

Ann Copestake, Peter Corbett, Peter Murray-Rust, C. J. Rupp, Advaith Siddharthan, Simone Teufel and Ben Waldron. 2006. An Architecture for Language Technology for Processing Scientific Texts. In Proceedings of the 4th UK E-Science All Hands Meeting. Nottingham, UK.

Peter Corbett, Colin Batchelor and Simone Teufel. 2007. Annotation of Chemical Named Entities. In Proceedings of BioNLP in ACL (BioNLP'07).

Peter T. Corbett and Peter Murray-Rust. 2006. High-throughput identification of chemistry in life science texts. *LNCS*, 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 Research*, Database Summary Paper 646.

The Gene Ontology Consortium. 2000. Gene Ontology: Tool for the Unification of Biology *Nature Genetics*, 25:25–29.

Ralph Delfs, Andreas Doms, Alexander Kozlenkov and Michael Schroeder. 2004. GoPubMed: Exploring PubMed with the GeneOntology. Proceedings of German Bioinformatics Conference, 169–178.

Christopher J. Mungall. 2004. Obol: integrating language and meaning in bio-ontologies. *Comparative and Functional Genomics*, 5:509–520.

Dietrich Rebholz-Schuhmann, Harald Kirsch, Miguel Arregui, Sylvain Gaudan, Mark Riethoven and Peter Stoehr. 2007. EBIMed—text crunching to gather facts for proteins from Medline. *Bioinformatics*, 23(2):e237–e244.

C. J. Rupp, Ann Copestake, Simone Teufel and Benjamin 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.