Workshop Programme

9:30 – 10:30 Invited talk (chair: Monica Monachini)
Mark Liberman, University of Pennsylvania & Linguistic Data Consortium

10:30 – 11:00 Coffee break

11:00 – 12:40 Session 1 (chair: Jian Su)

11:00 A Comparison of Knowledge Resource Designs: Supporting Term-level Text Annotation
A. Tribble, J. Kim, T. Ohta, J. Tsujii

11:30 The ITI TXM Corpora: Tissue Expressions and Protein-Protein Interactions
B. Alex, C. Grover, B. Haddow, M. Kabadjov, E. Klein, M. Matthews, S. Roebuck, R. Tobin, X. Wang

12:00 Semantic Annotation of Clinical Text: The CLEF Corpus
A. Roberts, R. Gaizauskas, M. Hepple, G. Demetriou, Y. Guo, A. Setzer, I. Roberts

12:20 Categorising Modality in Biomedical Texts
P. Thompson, G. Venturi, J. McNaught, S. Montemagni, S. Ananiadou

12:40 – 14:20 Lunch Break

14:20 – 16:00 Session 2 (chair: Goran Nenadic)

14:20 Static Dictionary Features for Term Polysemy Identification
P. Pezik, A. Jimeno, V. Lee, D. Rebholz-Schuhmann

14:50 Pyridines, Pyridine and Pyridine Rings: Disambiguating Chemical Named Entities
P. Corbett, C. Batchelor, A. Copestake

15:20 Chemical Names: Terminological Resources and Corpora Annotation
C. Kolářik, R. Klinger, C. Friedrich, M. Hofmann-Apitius, J. Fluck

15:40 Towards a Human Anatomy Data Set for Query Pattern Mining based on Wikipedia and Domain Semantic Resources
P. Wennerberg, P. Buitelaar, S. Zillner

16:00 – 16:10 Concluding remarks (chair: Sophia Ananiadou)

16:10 – 16:30 Coffee

Workshop Organisers

- Sophia Ananiadou, NaCTeM, University of Manchester, UK
- Monica Monachini, Istituto di Linguistica Computazionale, Pisa, Italy
- Goran Nenadic, University of Manchester, UK
- Jian Su, Institute for Infocomm Research, Singapore

Workshop Programme Committee

- Olivier Bodenreider, NLM, USA
- Paul Buitelaar, DFKI, Germany
- Nicoletta Calzolari, CNR, Italy
- Kevin B. Cohen, MITRE, USA
- Nigel Collier, National Institute for Informatics, Japan
- Walter Daelemans, University of Antwerp, Belgium
- Beatrice Daille, University of Nantes, France
- Udo Hahn, Jena University, Germany
- Marti Hearst, Berkeley, USA
- Martin Krallinger, Protein Design group, Spain
- Ewan Klein, Edinburgh University, UK
- Mark Liberman, CIS, UPenn, USA
- Hong Fang Liu, Georgetown University Medical Center, USA
- John McNaught, University of Manchester, UK
- Simonetta Montemagni, CNR, Italy
- Adeline Nazarenko, LIPN, Paris 13, France
- Claire Nedellec, CNRS, Framce
- John Pestian, Computational Medicine Center, Cincinnati Children's, USA
- Dietrich Rebholz-Schuhmann, EMBL-EBI, UK
- Patrick Ruch, University Hospital Geneva, Swiss Federal Institute of Technology
- Guergana Savova, Mayo Clinic, USA
- Hagit Shatkay, Queen's University, USA
- Stefan Schulz, Freiburg University Hospital, Germany
- Jun-ichi Tsujii, University of Tokyo, Japan and University of Manchester, UK
- Yoshimasa Tsuruoka, University of Manchester, UK
- Karin Verspoor, Los Alamos National Labs, USA
- Pierre Zweigenbaum, LIMSI-CNRS, France

Table of Contents

Author Index

FOREWORD

There has been tremendous work in biomedical text mining over the last decade. The size and coverage of the available literature and demands for text mining applications in the domains of biology and biomedicine are constantly increasing. These domains have become one of the driving application areas for the natural language processing community, resulting in a series of workshops and conferences that have reported on the progress in the field. Most of the work has focused on solving specific problems, often using task-tailored and private data sets. This data is rarely reused, in particular outside the efforts of the providers. This has changed during the last years, as a number of projects, initiatives and organisations have been dedicated to building and providing biomedical text mining resources (e.g. GENIA, PennBioIE, TREC Genomics track, BioCreative, Yapex, LLL05, BOOTStrep, JNLPBA, KDD data, Medstract, BioText, etc.). Although several resources have been provided for and from the community to support both training and evaluation of text mining applications, there have been few efforts to provide community-wide discussions on design, availability and interoperability of resources for bio-text mining.

The aim of this Workshop is to focus on building and evaluating resources used to facilitate biomedical text mining, including their design, update, delivery, quality assessment, evaluation and dissemination. Key resources of interest are lexical and knowledge repositories (controlled vocabularies, terminologies, thesauri, ontologies) and annotated corpora, including both task-specific resources and repositories reengineered from biomedical or general language resources. Of particular interest is the process of building annotated resources, including designing guidelines and annotation schemas (aiming at both syntactic and semantic interoperability) and relying on language engineering standards. Challenging aspects are updates and evolution management of resources, as well as their documentation, dissemination and evaluation.

The presented workshop papers cover many important aspects of biomedical resources. Several papers discuss features, design principles and experience in building lexical, terminological and knowledge resources, and present how these can be used to support different tasks, including term-level text annotations and disambiguation of different semantic classes (e.g. protein and gene names, chemical names and compounds, etc.). Also, the evolution of resources in a changing environment has been discussed, as well as using existing open sources (such as Wikipedia and domain semantic resources) to build terminologies and relation repositories. Building annotated corpora with different levels of terminological and functional mark-up has been of particular interest. Several papers present experience in building various biological, chemical and clinical corpora annotated with a range of entities and relations, including protein-protein interactions, tissue expressions, temporal relations and modalities. Quality of annotations (typically assessed through inter-annotator agreements) and future challenges (e.g. normalisation of entity mentions or mapping of relations) have been also widely discussed, proving that building useful and effective resources constitutes a major task in future biomedical text mining research.

The organisers would like to thank the authors for their valuable contributions and to the Program Committee members for their efforts in reviewing the submissions within a tight time frame. We are also grateful to the LREC 2008 organisers for logistical support and for hosting the event, and to the UK National Centre for Text Mining for sponsoring the workshop.

Sophia Ananiadou, Monica Monachini, Goran Nenadic, Jian Su

A comparison of knowledge resource designs: supporting term-level text annotation

Alicia Tribble, Jin-Dong Kim, Tomoko Ohta, Jun'ichi Tsujii
Department of Computer Science, University of Tokyo
Hongo 7-3-1, Bunkyo-ku, Tokyo 113-0033 JAPAN
{alicia, jdkim, okap, tsujii}@is.s.u-tokyo.ac.jp

Abstract

What makes a knowledge resource, like a domain model, thesaurus, or ontology, effective for term-level text annotation in the Biology domain? In this work we compare several approaches to ontology design with examples from well-known resources such as OBO, MeSH, and Genia. Based on these comparisons we establish goals for a knowledge resource that supports term-level text annotation and text mining: such a resource should represent terms and relations that are expressed in contiguous spans of text, and its terms should bear meaningful correspondences with other knowledge resources. Finally, we trace how these two goals have affected the re-design of the Genia Ontology over several iterations. The result is a new term hierarchy and a new design process, both specifically tailored to term-level text annotation. This research explores practical influences on the design of knowledge resources for Bio-NLP systems.

1. Introduction

A growing number of structured knowledge resources in the biology domain are providing guidance to researchers who perform large-scale data annotations for text mining and other language technologies. Some well-known examples include MeSH[1], Genia[2], and the ontologies of the Open Biomedical Ontologies Foundry[3](OBO), among many others.

For the purposes of this paper we refer loosely to such resources as "ontologies", although most are not proper ontologies in the philosophical sense. In fact, their structural differences reflect differences in their functional goals and in their appropriateness for certain NLP tasks. This paper explores those differences with the goal of identifying some of the important design properties of knowledge resources that support term-level annotation and text mining, specifically.

For example, consider the *biological process* ontology of the OBO Framework. OBO ontologies are designed "to be interoperable and logically well formed and to incorporate accurate representations of biological reality" (Smith, Ashburner et al. 2007). Their primary purpose is to serve as an accurate and standardized domain model. As a result, the biological process ontology divides cellular events into fine sub-categories such as *transcription* and its children via the IS-A relation: *RNA-dependent transcription (reverse transcription)* and *DNA-dependent transcription*.

These events are also represented in another large-scale knowledge resource: the Medical Subject Headings (MeSH), developed by the United States National Institutes of Health. MeSH organizes 97,000 terms ("entry terms") into a hierarchy of descriptors, where a link from child to parent descriptor denotes the relation "narrower-topic-than". In MeSH, the topic *Transcription* ambiguously covers both the generic sense of the term and the specific sense of *DNA-dependent transcription*. *Reverse Transcription* is the only child of this term. Justification for such an ambiguous structure can be derived from the context in which the MeSH hierarchy is used: MeSH terms serve to label and retrieve scientific papers from the PubMed/Medline database. Consider a user searching for papers related to "transcription". Taking DNA-dependent transcription as the most common meaning of this search term, a retrieval system has no need to distinguish between the generic and the specific. The MeSH hierarchy conserves this ambiguity, which corresponds to the way users search for documents, but still allows "reverse transcription" to be distinguished if necessary.

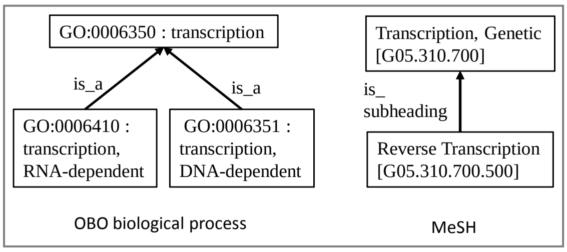

GO:0006350 : transcription
is_a
is_a
GO:0006410 : transcription, RNA-dependent
GO:0006351 : transcription, DNA-dependent
OBO biological process
Transcription, Genetic [G05.310.700]
is_ subheading
Reverse Transcription [G05.310.700.500]
MeSH

Figure 1. Partial hierarchies representing the biological event *Transcription* in OBO and MeSH

The difference between these two structures is shown in Figure 1. This is just one example of how well-researched, authoritative knowledge resources in a similar domain can differ in their design choices.

In this paper we compare three structured resources: MeSH, Gene Ontology (GO), and Genia version 1.0, showing how the intended use of each resource drives distinctions among them. Our goal is to identify properties of knowledge resources that support term-level annotation. *Term-level annotation* is the process by which human annotators classify the semantically significant expressions that present themselves in contiguous spans of text within a sentence, using the

[1] http://www.nlm.nih.gov/mesh/meshhome.html
[2] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/home/wiki.cgi
[3] http://www.bioontology.org/repositories.html#obo

terms from a structured knowledge resource as the annotation vocabulary. Entities and events that are expressed in passages longer than one sentence are outside the scope of term-level annotation. This type of annotated text is crucial for developing NLP tools including recognizers for named entities, biological events, and semantic relations. Without these tools, it would be difficult to imagine a system that could provide answers for a detailed biological query, such as "What are all the suppressors of MAP phosphorylation?" (Miyao, Ohta et al. 2006).

We characterize knowledge resources according to the granularity of their representational units and the interoperability of each resource with other knowledge resources in the domain. Through this process we arrive at design goals for resources that support term-level text annotation:

Granularity: *A structured knowledge resource for term-level annotation should represent terms and relations that are commonly expressed in contiguous spans of text in the target domain.*

Interoperability: *A structured knowledge resource for term-level annotation should represent consistently-defined relations among terms that bear meaningful correspondences with other authoritative knowledge resources in the target domain.*

Finally, we explore how the Genia Ontology in particular has improved on each of these goals over several design iterations (Genia version 1.0, Genia version 2.0, and Genia version 2.1). The Genia Ontology plays an important role as the annotation vocabulary for the Genia Corpus, a large-scale corpus of scientific abstracts that are annotated at the term level. Because of this role, its structure has been discussed and revised both internally and externally (Schulz, Beißwanger et al. 2006). By framing the evolution of the Genia Ontology in terms of granularity and interoperability, we can see how successive versions have improved in their fitness to support high-quality text annotation at the term level, which in turn leads to better-performing NLP systems in the biology domain. Such an analysis is useful to researchers faced with building or choosing knowledge resources for a biomedical NLP system.

1.2 Related work

Much work has been done in the domain of engineering and integrating formal ontologies in the biology domain. Smith, et al. (2004) present principles for formal classification and definition using the Gene Ontology as a subject of constructive critique. These principles are taken from best-practices in general ontology design and include *univocity*, *positivity*, and *objectivity*, among others. Zhang & Bodenreider (2006) present a similar examination of the Foundational Model of Anatomy[4], resulting in a different set of 15 design principles that cover *hierarchical cycles*, *incompatible relationships*, *implicit relations*, and the like.

These papers give valuable analysis of knowledge resources in terms of their formal consistency as stand-alone representational artifacts. Indeed, there are benefits that come from adopting a proper ontology as a biological domain model (machine-readability, automatic error detection in the model, logical inference, and others). However in the current paper we focus in more detail on the relationship between the design of ontologies (or ontology-like resources) and the annotation tasks they support. This line of reasoning follows in the tradition of (Tsujii and Ananiadou 2005), where the authors compare logical ontologies to thesauri in a text mining framework.

A similar task-oriented approach is used by Witte, Kappler et al. (2007), who present some principles for formal ontology design with a view to text mining and other natural language processing (NLP) tasks. These principles are expressed as component requirements for an ontology-driven text understanding system, including the following: *domain model*, *text model*, *biological entities*, and *entity relations*. A fundamental assumption of that work is that the semantic vocabulary used for text mining should be the same as the vocabulary used to perform higher-level reasoning over the results. Hence, their design principles are intended to result in formal ontologies that are appropriate for constraining the input to an automatic reasoner in the style of RACER (Haarslev and Moller, 2001).

A survey of ontological, terminological, and lexical resources for text mining is given in (Bodenreider 2006). The profusion of resources described in that work indicates the importance of taking a particular task, like term-level annotation, into account when choosing or building an effective knowledge resource.

Most of the principles described above govern structural properties and ontological soundness. These principles can be further described as contributing to the goal of effective ontology integration, and hence to the *interoperability* of the resulting ontologies. In developing our own principles, we are interested in both interoperability and *granularity*.

Granularity plays an important role in characterizing ontologies as NLP resources, particularly in the domain of biomedical text. The reason is that authors of scientific papers often use underspecified language to describe their results. They depend on the reader's contextual and background knowledge to interpret the precise meaning of words and expressions. Research in term-level annotation has shown that for consistency, annotated terms and relations must have concise evidence in text (Kim, Ohta et al. 2007). Hence, an ontology that supports such annotation, and as a result empirical NLP systems (for text mining, named-entity and event recognition, etc.) should reflect the granularity of language used in the target domain.

2. A comparison of ontologies based on intended use

Defining an appropriate level of interoperability and granularity for a knowledge resource depends on the target use of that resource. In this section we compare three existing biomedical ontologies according to their intended use: realistic domain modeling; classification, either of textual documents or of laboratory data (results

[4] http://fma.biostr.washington.edu/

<table>
<tr><th colspan="2">Covalent binding between N-acetyl-L-cysteine (NAC) and albumin was evaluated kinetically by conducting in vitro experiments.</th></tr>
<tr><td>Genia Annotation</td><td>Covalent <term sem=”binding”> binding </term> between
<term sem=”Amino_acid_monomer”> N-acetyl-L-cysteine (NAC) </term> and
<term sem=”Protein_molecule”> albumin </term> was evaluated kinetically by conducting in vitro experiments.</td></tr>
<tr><td>GO Annotation</td><td>GO ID: 0008144; Alb; CHEBI:28939; PMID:12458670; IPI</td></tr>
</table>

Table 1. Sample Annotations. The sentence shown is an excerpt from PubMed document PMID:12458670. Genia annotation is shown in-line in XML format. GO annotation is shown as a database entry that can be paraphrased as *drug binding, GO term 0008144, occurs between Alb and the chemical with CHEBI ID:28939, based on experimental evidence type IPI described in the paper with PubMed ID 12458670*.

Genia Leaf Term	**Appearance in contiguous spans of text**	**MeSH descendants**
Cells_cultured	“untransfected cell” “wild type cell line” “various cell line”	Cell line; 3T3 cells; Swiss 3T3 cells
Amino_acid_monomer	“acidic residue” “new amino acid” “CYS” “Nacetylcysteine”	Aminobutyric acids; gamma-Aminobutyric acids; Vigabatrin
Inorganic_compound	“physiological oxidant” “radical” “messenger molecule”	Alkalies; Carbonates; Lithium Carbonate

Table 2. Leaf terms from the Genia ontology often correspond to subtrees in MeSH. Examples of Genia terms in text, along with specializations that can be found under the corresponding term in MeSH.

databases); or annotating text at the term level. As a result, we can derive specific definitions that can be used in the context of each of these tasks.

2.1 Domain models: the Gene Ontology

The Gene Ontology, or GO (Consortium 2000) is a member of the Open Biomedical Ontologies Framework. It includes three sub-ontologies: *cellular component*, *molecular function*, and *biological process*, which was introduced earlier. Its purpose has evolved over time, and it currently serves as a domain model that represents terms and events grounded in biological truth.

As such, it is successfully used as a controlled vocabulary to label biology databases, so that experimental results from diverse laboratories can be automatically compared (Camon, Magrane et al. 2004). This annotation process relies on expert background knowledge as well as evidence in a broad sense, which comes from multiple documents in text, database, or other forms. GO annotations are described in detail on the GO website[5]. A sample GO annotation is given in Table 1.

The online GO FAQ[6] describes the annotation principles used in this process, as well as listing some target uses of the annotated data:

> *…first, every annotation must be attributed to a source, which may be a literature reference, another database or a computational analysis; second, the annotation must indicate what kind of evidence is found in the cited source to support the association between the gene product and the GO term.*
>
> *…applications for which GO has already been used include the following:*
> - *integrating proteomic information from different organisms;*
> - *assigning functions to protein domains;*
> - *verifying models of genetic, metabolic and product interaction networks.*

Based on this description, the observation can be made that GO annotation is performed roughly on the document level (where a document may be a summary of a laboratory result in a database). It allows the annotator to synthesize all of the evidence given by a single source. This has implications for the granularity of the ontology: terms are specific enough to support domain modeling on a fine scale, and they may represent entities or events that are only implicitly present in any single span of text.

That is, for a given document, the annotated ontology terms may not be explicitly represented by contiguous spans of text, but rather spread over the entire document. This feature of GO terms could help explain the results of (McCray, Browne et al. 2002), who searched automatically for GO term names in Medline text but found fewer matches than they expected.

GO annotations have also been used as part of the training and test data in the BioCreAtIvE Challenges. An investigation of the 2005 challenge results with respect to the task of aiding human annotation is given in (Camon, et al. 2005). An interesting result of that work was the comment by annotators that automatic protein labeling systems trained on GO data were less helpful than they

[5] http://www.geneontology.org/GO.annotation.shtml

[6] http://wiki.geneontology.org/index.php/GO_FAQ

could be as pre-processors because the systems returned long passages of text as evidence for candidate annotations. Human annotators preferred concise evidence, at most 5 lines in length, when judging whether to keep or modify an automatically-generated annotation. This result speaks to the role of fine-grained textual annotations, such as the term-level annotations we describe in this paper, in providing evidence and explanations of document-level classification results.

This GO usage scenario and the annotation principles also have an effect on the ontology's interoperability. Because they are intended to be shared across laboratories, GO terms are fairly precise and discriminative (although they do not list necessary and sufficient conditions for identifying terms).

The ontology conforms to OBO recommendations, and is distributed in multiple formats: OWL and OBO, among others. Relations between ontology concepts come from the OBO relations hierarchy (is-a, part-of), resulting in relatively clear & consistent relation definitions throughout GO (with some exceptions noted by (Smith, Köhler et al. 2004)). All of these features contribute to ease of understanding for researchers from diverse backgrounds who want to use GO to annotate their own experimental results.

2.2 Document classification hierarchies: MeSH

The MeSH term hierarchy was introduced in Section 1. Nelson and co-authors describe the goal of the hierarchy as follows:

> *…to provide a reproducible partition of concepts relevant to biomedicine for purposes of organization of medical knowledge and information.* (Nelson, Johnston et al. 2001)

In practice, MeSH terms are used to organize knowledge through the process of annotating scientific papers from the PubMed/MEDLINE[7] database. As documents are annotated with relevant concepts from the hierarchy, the documents themselves can be sorted, collected, organized, and retrieved more effectively.

Like GO annotations, MeSH annotations are made at the document level. However there are important differences. In GO, the referents of ontology terms are real-world biological entities; scientific papers are used as supporting evidence for applying a term to an instance of an entity in an experiment. Scientific knowledge of the entity increases as a result.

In MeSH, labels are applied directly to documents, and it is knowledge of the document that increases: i.e. what topic classes are dealt with in the document, what are the appropriate sub-headings, etc.

These differences have a crucial effect on the granularity of concepts that are appropriate for inclusion in MeSH. This topic is also addressed by Nelson, et al. (2001), who give the following example:

> *…MeSH contains a descriptor for 'Whales' but the domain of MeSH is biomedicine and not zoology. In the MEDLINE citation database, there are not sufficient citations to create a separate descriptor for each specific whale species. Nevertheless, it is useful to have the species names as entry terms to the descriptor. Gains in precision of retrieval by creating more specific descriptors would be small.*

Again, we see that the granularity of ontology terms (or in this case, terms from a hierarchical thesaurus) is derived from the intended use of the ontology. In MeSH this principle is applied in support of document retrieval by representing terms that "become important in conceptually partitioning the literature" (Nelson, Johnston et al. 2001).

The class and relation definitions in MeSH are more context-sensitive, and hence less interoperable, than in GO. However MeSH does support interoperability through detailed mappings between MeSH concepts and concept identifiers from comparable resources. For example, many major headings for chemicals in MeSH are mapped the corresponding structural name assigned by the Chemical Abstracts Service (CAS)[8] , a registry of over 33 million organic and inorganic substances.

Leaf-level mappings are an appropriate way of implementing interoperability for certain kinds of knowledge resources.

Because the MeSH hierarchy represents topical relations among its classes instead of biological relations, it would be difficult to map entire subtrees or graphs from MeSH onto a domain model like GO. In spite of this, identifying synonymous terms can certainly help users who are familiar with one ontology to understand the other more easily.

2.3 Structured vocabularies: Genia version 1.0

So far we have discussed the role of granularity and interoperability in domain models and in document classification hierarchies, using GO and MeSH as examples. These principles work in concert with the target use of the ontology to drive concrete design choices. In this section we turn to the Genia ontology, version 1.0[9], a knowledge resource that was designed specifically for text annotation at the term level.

Although "ontology" is part of its name, the Genia ontology version 1.0 is better described as a controlled vocabulary with a single-inheritance hierarchical structure. It includes 47 representational units that refer to biological continuants (non-event entities). Occurrents (events) are outside the scope of version1.0.

The design of the Genia ontology, version 1.0 is a direct response to the demands of term-level annotation in the Genia Corpus. Annotations are made as in-line XML markup to scientific abstracts that have been sampled from MEDLINE. An example is shown in Table 1. The latest release of the corpus includes 18,545 annotated sentences.

In designing the ontology, biologists familiar with the documents in the corpus selected a vocabulary of biological entities and roles that appear often enough in text to be consistently annotated. Next, a hierarchical ordering was imposed that places more general terms at the top of the hierarchy. Although the terms refer to

[7] http://www.nlm.nih.gov/pubs/factsheets/medline.html

[8] http://www.cas.org/expertise/cascontent/registry/

[9] http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/genia-ontology.html

biological entities, the hierarchy treats these terms more like topics, where a parent-child link indicates a "specialization-of" relation. Siblings in the hierarchy stand in a "topically-related" relation to one another. This allows annotators to find terms like *Protein_molecule* and *Protein_family_or_group* grouped under the common parent *Protein*. In a domain model, *Protein_family_or_group* might be moved to a branch representing sets. In the case of Genia, the "set" concept itself is outside the scope of annotation and hence left out of the hierarchy. As a result sets like *Protein_family_or_group* are placed near other terms to which they are topically related.

The Genia ontology, version 1.0 has a maximum depth of 6, more shallow than MeSH (depth 11) and GO (depth >= 7). This reflects the intent of the hierarchy to include only terms of a granularity that is pertinent to term-level annotation. Granularity is bounded at the most general by biological Substance and biological Source, dividing biological entities that can be described in terms of their chemical properties specific by terms that meet the standard for term-level annotation: these terms are commonly expressed in, and can easily be used to annotate, contiguous sub-sentential spans of text. Fine-grained distinctions that require additional context – a full document, as in MeSH, or full documents coupled with detailed background knowledge, as in GO - are considered beyond the scope of term-level annotation and as a result do not appear in the hierarchy.

This is a feature maintained in recent updates to the Genia ontology. Some examples from the Genia term ontology, version 2.0 are given in Table 2.

This table shows Genia leaf terms that correspond to internal nodes in the MeSH hierarchy. Typical examples of these terms appearing in contiguous spans of text are given in column 2. Column 3 shows how these classes are further refined in MeSH. Some strings bear the names of chemicals, but in many cases there is not enough evidence for sub-classification without additional context.

This structure imposes a low cost on annotators, whose goal is to quickly find the right term in the hierarchy for a textual expression that merits labeling. In addition, annotation principles that were developed to ensure high inter-annotator agreement have been translated into features of the ontology: terms that resulted in poor agreement were dropped from the vocabulary, and expressions that occurred often in text but fell outside the scope of the term hierarchy spawned additional terms. An example of a Genia term-level annotation is shown in Table 1.

3. Definitions of granularity and interoperability

The three resources described above demonstrate a pattern that links the granularity of concepts in a knowledge resource to the level on which annotation or retrieval is performed. Distinctions among biological entities can be made at finer levels of detail, given more contextual evidence.

Consider annotating the sentence: "I(kappa)B(beta) is constitutively phosphorylated.*"* Using the Genia term-annotation style, two entities can been annotated: a *Protein_molecule* ("I(kappa)B(beta)") and the process *Phosphorylation* ("phosphorylated"). As an alternative, by relaxing the requirement that annotations be assigned to contiguous spans of text, we could use the full sentence to determine a label The result is that a finer-grained subclass of Phosphorylation can be identified: *I-kappaB phosphorylation*. This subclass of phosphorylation is present in the Gene Ontology, where annotations are made at the document level.

This observation is supported by the granularity of terms we find in the Genia ontology, and can be expressed as a design goal for resources that support term-level annotation:

Granularity: *A structured knowledge resource for term-level annotation should represent terms and relations that are commonly expressed in contiguous spans of text in the target domain*.

Interoperability seems to be linked to annotation level more indirectly than granularity. In domain models like GO, interoperability is achieved by using individual term and relation definitions that are specific enough to be directly imported and exchanged among ontologies. In MeSH, sharing definitions and relations directly would impose class distinctions that conflict with the goal of appropriate granularity for document retrieval. The Transcription example given in Section 1.1 demonstrates how this can occur. Leaf-level mapping of some MeSH terms contributes to interoperability while allowing granularity to be the primary design goal.

In the Genia ontology, version 1.0 interoperability was not yet a design goal. However researchers both inside and outside the project have recognized the potential of clarifying the definitions of Genia terms and organizing them in a more consistent structure (Schulz, Beißwanger et al. 2006). This would increase interoperability and as a result annotations could become more accessible both to researchers and to downstream inference systems.

A working definition of interoperability at a level that supports term-level annotation should represent a commitment to making results interpretable, without imposing structural requirements that compete with the goal of granularity as it was just expressed. One such definition is the following:

Interoperability: *A structured knowledge resource for term-level annotation should represent consistently-defined relations among terms that bear meaningful correspondences with other authoritative knowledge resources in the target domain.*

4. Using design goals to drive ontology improvements

Genia, GO, and MeSH are all updated regularly in response to issues that arise during annotation. The first major redesign of the Genia ontology occurred when event annotation was added to the mission of the Genia project. A second revision effort is currently underway.

4.1 Genia 1.0 to Genia 2.0

In the first major revision of the ontology, designers have removed the Substance/Source distinction from the hierarchy, refined the class definitions, and added a new section of the ontology to cover biological occurrents. The Genia ontology, version 2.0 is a single-inheritance

Genia version 2.0 Term	Example sentence
Protein_molecule	"We have detected a specific nuclear protein complex that binds to the element and show that **NF-kappa B1** (p50) is a part of this complex." "In contrast, **NF-kappa B p50** alone fails to stimulate kappa B-directed transcription, and based on prior in vitro studies, is not directly regulated by I kappa B."
Protein_complex	"Analysis of the nuclear extracts with antibodies directed against the major components of **NF-kappa B** the p50 and RelA (p65) proteins, indicated that the composition of NF-kappa B was similar in neonatal and adult cells." "This was due to the presence of active **NF-kappa B complexes** in the nucleus of CD45- T cells."
Protein_domain_or_region	"Does nucleolin bind the NF kappa B DNA binding motif?"
Protein_family_or_group	"Besides p50, 1,25(OH)2D3 decreased the levels of another **NF-kappa B protein**, namely c-rel." "Fibrinogen activates **NF-kappa B transcription factors** in mononuclear phagocytes."

Table 3. Example annotations using subclasses of Protein from Genia 1.0/2.0. Annotated strings are shown in bold.

hierarchy represented in OWL-DL. It is included in a Genia corpus release that can be downloaded from the project homepage .

The first change, removing *Source* and *Substance*, results in a hierarchy where the criteria that distinguish one branch from another are less opaque. The hierarchy rooted at *Substance* originally referred to entities that could be classified according to chemical structure. In updated versions this sub-tree tree is headed by the more revealing term *Chemical*, but it maintains the same depth and sub-classes as the *Substance* hierarchy. The granularity of terms in this branch is unaffected although interoperability improves, since the new naming conventions more closely match those used in other ontologies.

The term *Source* has been replaced with two subtrees, *Anatomy* and *Organism*. This allows the vacuous distinction between *Natural_source* and *Artifical_source*, which never appeared in textual annotations, to be replaced with natural/cultured distinctions on the frontier of the hierarchy, where they do appear in text. Some examples are given in Table 4. This change improves the granularity of the affected classes.

Many of the class definitions in version 1.0 of the Genia ontology were inductive, providing example members and depending on the reader to infer a definition. The revised class definitions are heavily borrowed from definitions of MeSH terms. The change to declarative definitions makes the criteria for membership in Genia classes more clear. In addition, domain experts carefully performed a mapping from Genia classes to MeSH terms during this process. The small size of the Genia ontology (46 classes) makes this a task appropriate for human annotators, rather than machine learning tools. As a result, meaningful correspondences between Genia and MeSH were identified, increasing interoperability of the Genia ontology.

4.2 Genia 2.0 to Genia 2.1

The designers of Genia have continued to refine the ontology since releasing version 2.0 in (Kim, Ohta et al. 2007). Current changes being considered include redefining the parent-child relation among classes and moving terms that refer to textual features or roles into a separate hierarchy.

Section 2.3 describes the parent-child relationship in Genia version 1.0 as "specialization-of", and this relation holds in version 2.0, as well. This arrangement is most clearly seen in the subclasses of *Protein*. In text, the same protein name can be used to refer to entities of very different types: sets of proteins (Genia class *Protein_family_or_group*), parts of proteins (Genia classes *Protein_substructure* and *Protein_domain_or_region*), and subclasses of proteins (Genia classes *Protein_molecule* and *Protein_complex*). Table 3 presents some examples.

Although they do not represent entities that stand in a biological *is-a* relation, placing these terms under *Protein* in the Genia ontology, version 2.0 reflects the fact that, when none of these specializations applies, annotators should use the more general parent class to annotate a protein name in text.

In the Genia term ontology, version 2.1, subclasses like *Protein_family_or_group* have been removed to a separate hierarchy of *Expression_features*. The new relation *has_text_feature* is defined between a Genia term and an expression feature. Using this relation, we can still create an annotation for a protein name that is used in text to refer to a family or group, but the connection between the protein and the family is clearly indicated as a product of textual usage, rather than biological inheritance. This also applies to the subclasses of *DNA* and *RNA* from version 2.0 of the ontology. The updated ontology is shown in Figure 2.

Annotation now occurs at the same level of granularity as before, using this combination of Genia term and expression feature. With *Protein_family_or_group* and its siblings removed, the remaining classes in the Genia term ontology, version 2.1 refer to biological entities that stand in the traditional *is-a* relation to each other. This structural consistency improves interoperability, according to our definition.

The new structure also gives us a fresh perspective on how expression features interact with the biological entities. We can observe that expression features are currently applied to instances of *DNA*, *RNA*, and *Protein*.

Cell_cultured	“resting Jurkat cell” “various cell line”
Cell_natural	“B lymphocyte” “APL blast”
Tissue_cultured	“fetal thymic organ culture”
Tissue_natural	“airway tissue” “endothelium”

Table 4. Example strings annotated with new classes from Genia 2.0

The structure of the ontology suggests that there may be a more elegant compositional solution: allow all instances of *Organic_compound* to be modified by expression features. In addition, the expression features themselves could be regrouped or subclassed to improve consistency. This type of compositional analysis, in the tradition of (Ogren, 2005), is the subject of our ongoing work on refining the Genia ontology.

5. Conclusion

The design strategies used in the most recent revisions of the Genia ontology are aimed at improving its interoperability while maintaining a level of granularity that supports term-level annotation of biological entities. A comparison of MeSH, GO, and Genia revealed that these features can be used to characterize and compare structured knowledge resources, and that design choices can be motivated directly by the intended use of the resource. All three of these resources are the object of ongoing development and research. Future releases of the Genia ontology will continue to bear these lessons in mind.

6. Acknowledgements

This work was partially supported by Grant-in-Aid for Specially Promoted Research (Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan), and Genome Network Project (MEXT, Japan).

7. References

Bodenreider, O. (2006). Lexical, terminological and ontological resources for biological text mining. Text mining for biology and biomedicine. S. Ananiadou and J. McNaught, Artech House: 43-66.

Camon, E. B., D. G. Barrell, et al. (2005). "An evaluation of GO annotation retrieval for BioCreAtIvE andGOA." BMC Bioinformatics 6(Suppl 1)(S17).

Camon, E., M. Magrane, et al. (2004). "The Gene Ontology Annotation (GOA) Database: sharing knowledge in Uniprot with Gene Ontology." Nucleic Acids Res 32(Database Issue)(D262-D266).

Consortium, T. G. O. (2000). "Gene Ontology: tool for the unification of biology." Nature Genet 25: 25-29.

Haarslev, V. and R. Moller (2001). RACER System Description. International Joint Conference on Automated Reasoning (IJCAR), Siena, Italy, Springer-Verlag.

Kim, J.-D., T. Ohta, et al. (2007). "Corpus annotation for mining biomedical events from literature." BMC Bioinformatics 9(10).

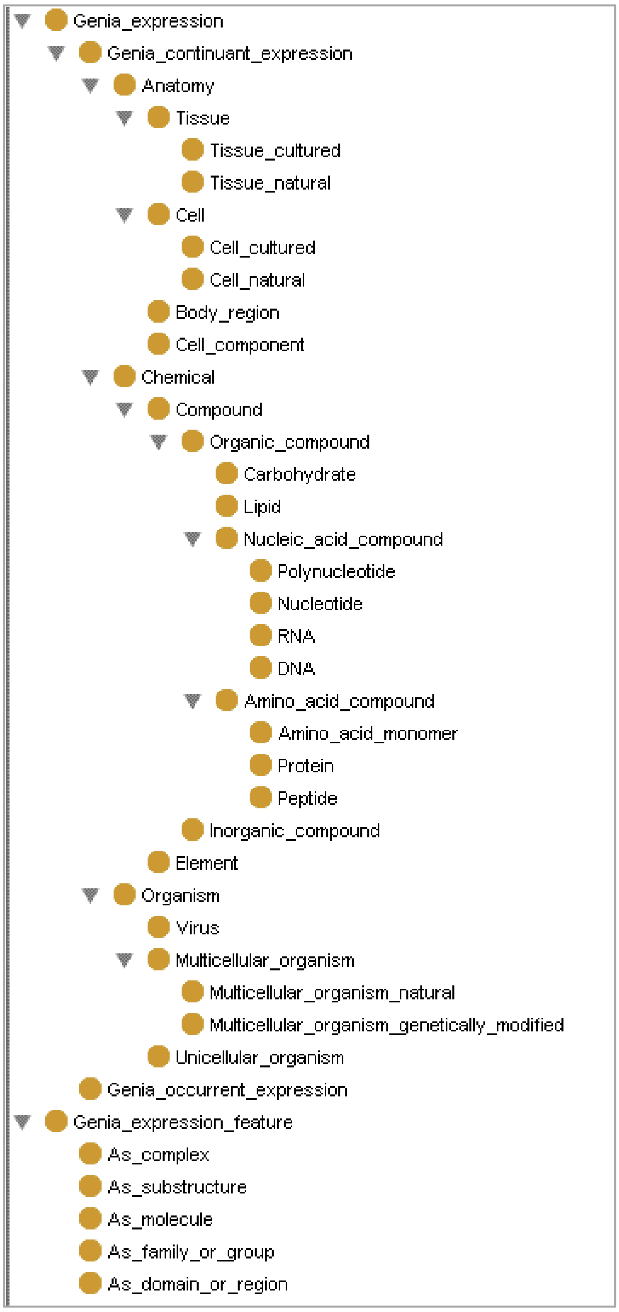


Figure 2. Genia Term Ontology, version 2.1 (proposed)

Miyao, Y., T. Ohta, et al. (2006). Semantic Retrieval for the Accurate Identification of Relational Concepts in Massive Textbases. COLING-ACL 2006, Sydney, Australia.

Nelson, S. J., D. Johnston, et al. (2001). Relationships in Medical Subject Headings. Relationships in the organization of knowledge. C. A. Bean and R. Green. New York, Kluwer Academic Publishers: 171-184.

Ogren, P. V., K. B. Cohen, et al. (2005). Implications of Compositionality in the Gene Ontology for Its Curation and Usage. Pacific Symposium on Biocomputing.

Schulz, S., E. Beißwanger, et al. (2006). From GENIA to BioTop - Towards a top-level Ontology for Biology. International Conference on Formal Ontology in Information Systems (FOIS 2006), Baltimore, USA.

Smith, B., M. Ashburner, et al. (2007). "The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration." Nature Biotechnology 25: 1251 - 1255.

Smith, B., J. Köhler, et al. (2004). On the Application of Formal Principles to Life Science Data: A Case Study in the Gene Ontology. Database Integration in the Life Sciences (DILS 2004), Berlin.

Tribble, A. and C. Rosé (2006). Usable Browsers for Knowledge Acquisition. Poster Presentations of CHI-2006. Montreal, Quebec.

Tsujii, J.-i. and S. Ananiadou (2005). "Thesaurus or logical ontology, which do we need for mining text?" Language Resources and Evaluation [Journal of LRE], Springer SBM 39(1): 77-90.

Witte, R., T. Kappler, et al. (2007). Ontology Design for Biomedical Text Mining. Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences. C. J. O. Baker and K.-H. Cheung. New York, NY, USA, Springer Science+Business Media: 281--313.

Zhang, S. and O. Bodenreider (2006). "Law and order: Assessing and enforcing compliance with ontological modeling principles in the Foundational Model of Anatomy." Computers in Biology and Medicine 36: 674–693.

The ITI TXM Corpora: Tissue Expressions and Protein-Protein Interactions

**Bea Alex, Claire Grover, Barry Haddow, Mijail Kabadjov,
Ewan Klein, Michael Matthews, Stuart Roebuck, Richard Tobin and Xinglong Wang**

University of Edinburgh, School of Informatics
2 Buccleuch Place, Edinburgh, EH8 9LW, Scotland, UK
txm-researchers@inf.ed.ac.uk

Abstract

We report on two large corpora of semantically annotated full-text biomedical research papers created in order to develop information extraction (IE) tools for the TXM project. Both corpora have been annotated with a range of entities (CellLine, Complex, DevelopmentalStage, Disease, DrugCompound, ExperimentalMethod, Fragment, Fusion, GOMOP, Gene, Modification, mRNAcDNA, Mutant, Protein, Tissue), normalisations of selected entities to the NCBI Taxonomy, RefSeq, EntrezGene, ChEBI and MeSH and enriched relations (protein-protein interactions, tissue expressions and fragment- or mutant-protein relations). While one corpus targets protein-protein interactions (PPIs), the focus of other is on tissue expressions (TEs). This paper describes the selected markables and the annotation process of the ITI TXM corpora, and provides a detailed breakdown of the inter-annotator agreement (IAA).

1 Introduction

This paper describes two corpora constructed and annotated for the TXM project. The aim of the TXM project was to develop tools for assisting in the curation of biomedical research papers. The ITI TXM corpora were used to train and test machine learning based NLP components which were interfaced with a curation tool.

There already exist several corpora of annotated biomedical texts (Section 2), all with individual design and annotation characteristics. The ITI TXM corpora combine a number of attractive characteristics of such available corpora, thus making them a valuable resource for NLP research. We annotated full-text papers since our intended target application (the curation tool) worked with such documents. Furthermore, it has been shown in previous research that there is valuable information in full-text articles that cannot be obtained from their abstracts alone (e.g. by Shah et al., 2003 and McIntosh & Curran, 2007). The markables used in the ITI TXM corpora included not only a range of named entities and relations, but also extensive, multi-species normalisation of proteins, genes and other entities, to standard publicly available databases.[1] Furthermore, some of the relations were enriched with additional biomedical information enabling finer-grained classification, and connecting the relations with other entities in the text. At around 200 full-text papers each, the corpora are relatively large in size. In addition, we will release multiple annotations of many of the papers, enabling the comparison of different annotators' views of the corpus. The set of markables chosen for both corpora arose out of extensive discussions between biologists managing the curation, and NLP researchers creating the NLP components. The biologists were consulted to determine what information they wanted to be extracted. At the same time, their ideas had to be balanced against what was possible using the state-of-the-art in NLP technology, and what could be reliably annotated. The final set of markables resulted out of several iterations of piloting and measurements of IAA.

This paper is organised as follows: after discussing related work on biomedical corpus design and annotation in the next section, a description of how the documents were selected for the corpora is provided in Section 3. An overview of both corpora, a description of the markables, the annotation process and details of the IAA are presented in full in Section 4. Finally Section 5 offers some conclusions and lessons learnt from the annotation project.

2 Related Work

In recent years, there have been numerous efforts in constructing and annotating biomedical corpora. Comprehensive lists of publicly available corpora are maintained by Cohen et al.[2] as well as Hakenberg[3]. This related work section does not provide an all-inclusive list of biomedical corpora but rather presents different characteristics of corpus design and annotation illustrated by typical examples.

Existing resources vary in size, type of data, markables and levels of annotation, the way the annotation is applied, their distributed formats and their domains. The GENIA corpus (Ohta et al., 2002), for example, is one of the largest and most widely used data sets in the text mining community. It consists of 2,000 Medline abstracts and is manually annotated with a series of semantic classes defined in the GENIA ontology. Other corpora are made up of sets of sentences from biomedical research articles, as is the case for BioInfer (Pyysalo et al., 2007) and GENETAG (Tanabe et al., 2005). The latter is a collection of 20,000 Medline sentences annotated for gene and protein names in one semantic class. Parts of this corpus were used in the BioCreAtIvE I and II competitions that, amongst other tasks, enabled different text mining research groups to evaluate how well their systems perform at extracting gene/protein names from biomedical literature.

Although there have been a series of corpus construction efforts for the purpose of biomedical text mining, only a small number of groups (e.g. Wilbur et al., 2006 and Krallinger et al., 2006) report IAA figures. In other words, it is rare to find information about how consistent two independent

[1] Normalisation refers to the task of grounding a biomedical term in text to a specific identifier in a referent database. See Table 3 for the publicly available databases used.

[2] http://compbio.uchsc.edu/ccp/corpora/obtaining.shtml

[3] http://www2.informatik.hu-berlin.de/~hakenber/links/benchmarks.html

annotators are when marking up a representative sample of a data set. The assumption is that the level of IAA provides insights into how challenging a particular task is to a human expert, providing an upper bound for an automated system is and how appropriate the task in itself is. Lu et al. (2006) show an increase in IAA over time as annotators become more familiar with their task of marking up GeneRIFs with 31 semantic classes in the protein transport domain. Figures of IAA also help to determine weaknesses in the annotation guidelines. Mani et al. (2005) measured IAA based on a first set of annotation guidelines for marking up protein names.[4] After analysing the annotation differences, they revised their guidelines which resulted in an improvement of IAA in a second annotation round and simultaneously in better annotation quality overall. Alex et al. (2006) have shown that consistency in the annotation of named entity boundaries is crucial to obtain high accuracy for biomedical named entity recognition. The need for both clear annotation guidelines to achieve such consistency and comprehensive annotation guidelines to capture complex information in unstructured text data is often highlighted (e.g. see Wilbur et al., 2006 and Piao et al., 2007). Making such guidelines available to the research community and publishing figures of IAA is recommended by Cohen et al. (2005) who analysed the characteristics of different biomedical corpora. They also conclude that distributing data in standard formats (e.g. XML) is vital to guarantee high corpus usage.

As mentioned earlier, publicly available corpora differ in the type of textual data, i.e. a corpus can be made up of sentences, abstracts or full-text papers. McIntosh & Curran (2007) and Shah et al. (2003) indicate a clear need for biological IE from full-text articles. The former study shows that only a small proportion of identified fact instances appears in abstracts. The latter found that although abstracts contain the best ratio of keywords, other sections of articles are a better source of biologically relevant data. As a result, they advocate IE systems that are tuned to specific sections. As much of the important information is not present in the abstract but the main paper, Cohen et al. (2005) suggest that abstracts and isolated sentences are inadequate and unsuited to the opportunities that are available for text mining. Sometimes, the most relevant information in a paper is found in figure captions (Shatkay and Feldman, 2003). Currently, only few available resources contain full-text publications, one example of such a corpus being FetchProt (2005). Its annotation includes specific experiments and results, the proteins involved in the experiments and related information. Exploiting such full-text resources is vital to develop text mining systems that will be used in practice, e.g. by biologists, clinicians or curators.

Publicly available biomedical corpora also often differ in their markables and levels of annotation. Some are annotated with part-of-speech tags (e.g. GENIA) and named entities, most often gene/protein names (e.g. GENETAG) that are sometimes normalised to identifiers (e.g. FetchProt). In other cases, the annotation includes binary relations between entities such as PPIs (e.g. AImed described in Bunescu et al., 2005) or non-binary relations (e.g. BioInfer). Several corpora are distributed with syntactic annotation such as phrase-based or dependency-based structures, e.g. BioIE (Kulick et al., 2004), GENIA treebank (2005), LLL (Nedellec, 2005) and BioInfer.

In this paper, we introduce two large biomedical corpora in the sub-domains of PPIs and TEs which will be distributed in one collection as the ITI TXM corpora. Both corpora are made up of full-text papers that are annotated with a series of relevant named entities, some of which are normalised. Furthermore, the annotations include various types of relations as well as relation attributes and properties (see Section 4.2). Domain experts used extensive curation guidelines that were devised based on several rounds of piloting (see Section 4.3). We provide figures of IAA for all types of semantic annotation for a representative corpus sample (see Section 4.4). Moreover, the data is distributed in XML with semantic annotations in standoff format (Carletta et al., 2005). In the future, the ITI TXM corpora will serve as a valuable resource to train IE methods for mining facts from biomedical literature.

3 Document Selection

Document selection for the PPI corpus was performed in two stages. The initial plan was to annotate only full-text articles available in XML. Therefore, 12,704 full-text XML files were downloaded from PubMedCentral OpenAccess.[5] The documents were filtered by selecting those articles that contained at least 1 of 13 terms either directly associated with PPIs or with biological concepts representative of typical curation tasks.[6] The abstracts and, if necessary, full texts of the remaining 7,720 documents were all examined by trained biologists and selected if they contained interactions that were experimentally proven within the paper, resulting in a total of 213 documents.[7] In order to ensure that enough documents were available for annotation, the same queries were performed against PubMed and additional documents were selected from the resulting list using the same criteria.[8] Several of the documents were excluded from the final set because they were used during the piloting or were rejected by the annotators as not being suitable for annotation. The resulting corpus consists of 217 documents, 133 selected from PubMedCentral and 84 documents selected from the whole of PubMed.

Document selection for the TE corpus was performed against PubMed. This was partially to ensure that enough documents were selected, and partially to address the concern that in practice, many important documents would not be available in XML and the annotations would be more representative if they accounted for this reality. The initial pool of documents was selected from PubMed using terms designed to capture documents representative of typical TE and PPI curation tasks.[9] The abstracts of the resulting 12,060 documents were randomised and examined

[4] Mani et al. (2005) refer to IAA as inter-coder reliability.

[5] `http://www.pubmedcentral.nih.gov/`
The 12,704 articles represented the complete set of available documents on 17/08/2005.

[6] The terms were: *bind*, *complex*, *interact*, *apoptosis*, *ubiquitination*, *mitosis*, *nuclear envelope*, *cell cycle*, *phosphorylation*, *glycosylation*, *signal transduction* and *nuclear receptors*.

[7] Clinical articles on drug or patient trials were excluded.

[8] `http://www.ncbi.nlm.nih.gov/PubMed/`

[9] The queries were: *"Gene Expression Regulation"[MeSH]*, *Development*, *"Signal Transduction"[MeSH]*, *"Protein Biosynthesis"[MeSH]*, *"Cell Differentiation"[MeSH]*, *Apoptosis*, *Mitosis*, *Cell cycle* and *Phosphorylation*

Annotations	PPI				TE			
	TRAIN	DEVTEST	TEST	All	TRAIN	DEVTEST	TEST	All
1	65	25	35	125	82	34	34	150
2	48	9	8	65	68	7	11	86
3	20	5	2	27	1	0	1	2
Total documents	133	39	45	**217**	151	41	46	**238**
Total annotations	221	58	57	**336**	221	48	59	**328**

Table 1: Counts of numbers of papers with 1, 2 or 3 annotations in each section of each corpus.

in order by a biologist and selected if they contained mentions of the presence or absence of mRNA or protein in any organism or tissue. A total of 4,327 documents were examined of which 1,600 were selected for TE annotation. The TE corpus is comprised of the first 238 of these documents that were not used during piloting and not rejected by the annotators.

In both phases, documents were split into TRAIN, DEVTEST, and TEST sets in a ratio of approximately 64:16:20 (see Table 1). TRAIN was to be used for training machine learning models and deriving rules, DEVTEST for testing during system development, and TEST for testing the final system. The document selection methods were dictated, in part, by the requirements of the industrial partner that assisted in the annotation of the corpora. The terms used were based on the queries used for selecting documents for creating commercially viable curated databases. Furthermore, the results of document selection were used to create training and testing corpora for a document retrieval system designed to improve the document selection phase. These corpora will be released at a future date.

4 Corpus Annotation

4.1 Overview

Documents were selected for annotation as described in Section 3. The full-text papers were downloaded from PubMed or PubMedCentral either as XML, or as HTML if the XML version was not available, and then converted to an in-house XML format using LT-XML2 tools.[10] The LT-XML2 and LT-TTT2 tools were also used to tokenise and insert sentence boundaries into the text (Grover et al., 2006).

From each corpus a random selection of documents was chosen for double or triple annotation in order to allow calculation of IAA, which is used to track annotation quality and to provide a measure of the difficulty of the task. The counts of singly and multiply annotated documents in the TRAIN, TEST and DEVTEST sections for both corpora are shown in Table 1. Multiply annotated documents were left in the corpus and not reconciled to produce a single, gold standard version. It was found during piloting that reconciliation could be very time-consuming so we decided to focus our resources on obtaining a larger sample of papers.

During the annotation of the full-text papers, we did not annotate sections that did not contain any relevant information, e.g. contact details and reference sections, HTML navigational text. Moreover, materials and methods sections were not annotated on the grounds that they would be too time-consuming to annotate. The annotators marked unannotated paragraphs during the annotation so that these sections could be excluded from training and testing. Based on the sentence splitting and tokenisation performed during the pre-processing, the PPI corpus contains approximately 74.6K sentences and 2.0M tokens, and the TE corpus is made up of around 62.8K sentences and 1.9M tokens.[11]

Entity type	PPI	TE
CellLine	7,676	—
Complex	7,668	4,033
DevelopmentalStage	—	1,754
Disease	—	2,432
DrugCompound	11,886	16,131
ExperimentalMethod	15,311	9,803
Fragment	13,412	4,466
Fusion	4,344	1,459
GOMOP	—	4,647
Gene	—	12,059
Modification	6,706	—
mRNAcDNA	—	8,446
Mutant	4,829	1,607
Protein	88,607	60,782
Tissue	—	36,029

Table 2: Entity types and counts in each corpus. A long dash indicates that the entity was not marked in that corpus.

4.2 Description of Markables

In both corpora the markables, i.e. units of annotation, consist of named entities, normalisations, relations, properties and attributes.

Named entities are terms of interest to biologists which belong to pre-defined semantic classes. Table 2 shows the named entity types marked and their counts in each corpus. In the PPI corpus, the entities are either proteins and other related entities involved in PPI relations (Protein, Complex, Fusion, Fragment and Mutant) or attributes of PPI relations (CellLine, DrugCompound, ExperimentalMethod, Modification). Conversely, for the TE corpus, the entities are either those that can be involved in TE relations (Tissue, Protein, Complex, Fusion, Fragment, Mutant, Gene, mRNAcDNA and GOMOP) or those that can be attributes of TE relations (DevelopmentalStage, Disease, DrugCompound, ExperimentalMethod). All named entity types (except GOMOP) have intuitively obvious biological interpretations, which are made precise in the annotation guidelines. For example, the definition of DrugCompound is: “a chemical substance of known composition used to affect the function of an organism, cell or biological process”. The GOMOP entity type was used in cases where the annotator felt that the author was referring to a “***Gene or mRNAcDNA or Protein***”. We felt that having a single entity type to represent this kind of ambiguity would be simpler than allowing annotators to mark the same term as multiple entity types (e.g. Protein and Gene).

[10] http://www.ltg.ed.ac.uk/software/xml/

[11] Note that all annotated versions of each paper are treated as separate documents in this calculation.

Database	Url	Prefix	PPI	TE
NCBI Taxonomy	http://www.ncbi.nlm.nih.gov/Taxonomy/	ncbitaxon:	Protein	Gene, mRNAcDNA, Protein, GOMOP
RefSeq	http://www.ncbi.nlm.nih.gov/RefSeq/	refseq:	Protein	Protein, mRNAcDNA
EntrezGene	http://www.ncbi.nlm.nih.gov/entrez/	gene:	Protein	Gene, mRNAcDNA, Protein, GOMOP
ChEBI	http://www.ebi.ac.uk/chebi/	chebi:	—	DrugCompound
MeSH	http://www.nlm.nih.gov/mesh/	mesh:	—	Tissue

Table 3: Databases used for normalisations and the entities to which they are assigned in each corpus. A long dash indicates that the database was not used in that corpus.

Corpus	Relation type	Count
PPI	PPI	11,523
PPI	FRAG	16,002
TE	TE	12,426
TE	CHILD-PARENT	4,735

Table 4: Relation types in each corpus.

When marking named entities, the annotators were permitted to nest them, but entities were not allowed to cross. For any pair of entities with a non-empty intersection, the intersection therefore had to coincide with at least one of the entities. Entities were also required to be continuous. Discontinuous coordinations such as “A and B cells” were annotated as two nesting entities “A and B cells” and “B cells”, indicating that the first was discontinuous using a flag in the XML. Furthermore, annotators were able to override the tokenisation if entity boundaries and token boundaries did not coincide, by indicating the entity boundaries using character offsets. For example, in one annotated document, the term “Cdt1(193-447)” is tokenised as a single token, but the annotator decided that “Cdt1” was a Protein and “193-447” was a Fragment. The Protein was therefore marked using an end offset of -9, to indicate that the end of the Protein name was 9 characters from the end of the token, and in a similar way the Fragment had start offset 5 and end offset -1. The XML representation of the data enables retokenisation as proposed by Grover et al. (2006) to improve the original tokenisation at a later stage while preserving the entity annotation.

A number of types of entities were normalised to one or more of the standard, publicly available biomedical databases listed in Table 3. In general, for each entity term that was normalised, an ID of the appropriate database was assigned as the normalisation value with a prefix indicating the source database. If no appropriate identifier existed, the ID was left blank and only the database prefix was used as the normalised value.

Normalisation of protein, gene and mRNAcDNA entities was more complex. Two types of normalisations were added to each occurrence of such entities: *full normalisation* and *species normalisation*, where the former involves assigning RefSeq identifiers to protein and mRNAcDNA terms and EntrezGene identifiers to gene terms; and the latter involves assigning NCBI taxonomy identifiers to protein, gene and mRNAcDNA terms. The project initially aimed at providing *full normalisation* for both corpora.[12] However, *full normalisation* turned out to be too time-consuming. Given limited time and resources, only the TE corpus and the DEVTEST and TEST portions of the PPI corpus were fully normalised, while the TRAIN portion of the PPI corpus was only species-normalised. A few special cases must be considered in the normalisation annotation:

- *Species mismatch.* For the term to be normalised, there is an entry in the database (e.g. RefSeq) which matches the specific entity but the entry does not match the species of the term given the surrounding context. In this case the term was only normalised for its species (i.e. species normalisation).
- *Several host species.* The term to be normalised is discussed relative to several host species. In this case, the term was normalised multiple times and each annotated entity was assigned a unique identifier for each species mentioned. In case of more than five possible host species for the term, annotators followed the next instruction.
- *Host species not clear.* The host species of a term to be normalised cannot be determined from the text, because it is discussed in a general way rather than in relation to one or more specific species, or the text is unclear about the host species of the term. In this case, the entity was normalised as if its species was *Homo sapiens*, and the keyword “gen” (for “general”) was added to any chosen identifier, e.g. “NP_004513 (gen)”, and at the same time the Taxonomy identifier for Homo sapiens together with the keyword “gen” (e.g., “9606 (gen)”) were entered as the species-normalisation. However, if *Homo sapiens* could not possibly be the correct host species, due to the occurrence of a general species word, such as *viral* or *bacterial*, “gen” was entered for species normalisation.

In each corpus, two types of relations were marked (see Table 4). In the PPI corpus, relations refer to interactions between two proteins (PPI) and connect Mutants and Fragments with their parent proteins (FRAG). In the TE corpus, relations indicate when a gene or gene product is expressed in a particular tissue (TE); relations also connect Mutants and Fragments with their parent proteins (CHILD-PARENT). Annotators were permitted to mark relations between entities in the same sentence (intra-sentential) and in different sentences (inter-sentential). For the TE and PPI relations, annotators also marked “link terms” used by the authors to indicate a relation. Marked in the same way as entities, these are called InteractionWord for PPI relations and ExpressionLevelWord for TE relations.

The properties and attributes are extra pieces of information added by the annotators to both PPI and TE relations. A property is a name-value pair assigned to a relation to add extra information, for example whether a PPI is mentioned

[12] In fact, both RefSeq and EntrezGene identifiers are species-specific. When a term is “fully normalised” its host species can therefore be identified without *species normalisation*.

Name	Value	PPI	TE
IsPositive	Positive	10,718	10,243
	Negative	836	2,067
IsDirect	Direct	7,599	—
	NotDirect	3,977	—
IsProven	Proven	7,562	9,694
	Referenced	2,894	1,837
	Unspecified	1,096	736

Table 5: Property names, values and counts in each corpus. A long dash indicates that the property was not marked in this corpus.

as being direct or indirect, or whether it was experimentally proven in the paper. Both positive and negative TE and PPI relations, i.e. statements asserting that an interaction or expression did or did not occur, were also marked, with properties used to distinguish between them. The names and values for the properties were drawn from a small closed list and annotators assigned at least one value to each name, for each relation. Their counts in each corpus are listed in Table 5.

Attributes are named links between relations and other entities, e.g. to indicate the experimental method used to verify a PPI relation, or the cell line used to discover a TE relation. In the PPI corpus, all attributes, except for MethodEntity, are attached to entities. Conversely, all attributes are attached to relations in the TE corpus. Attributes are also used to link a relation to its link term and do not have to be in the same sentence as the relation. The names and counts of the attributes are listed in Tables 6 and 7.

Note that as well as being able to add multiple values for each relation property, annotators were also permitted to add multiple values for each attribute. They did this by marking extra relation entries. For example, in a sentence such as “Protein A interacts with B in the presence of Drug C but not D.”, the annotators would mark two PPI relations between “A” and “B”, one Positive with “C” as a DrugCompound attribute, and the other negative with “D” as a DrugCompound attribute.

4.3 The Annotation Process

Annotation was performed by a group of nine biologists, all qualified to PhD level in biology, working under the supervision of an annotation manager (also a biologist) and collaborating with a team of NLP researchers. At the beginning of the annotation of each corpus, a series of discussions between the biologists and the NLP team were held with the aim of determining a set of markables. Since the overall aim of the project was to build NLP tools for integration into a curation assistant, the markables suggested by the biologists were those which they wished the curation assistant to aid them with. The NLP team provided input as to which markables might be technically feasible and what could be reasonably accomplished within the project timescale.

A further consideration in selecting markables was how well they could be annotated in practice. Markables which could not be reliably annotated by humans would not produce good data, and as a result would be even more difficult for automated systems to extract. Using the initial list of markables, several rounds of piloting were conducted to determine the markables that could be annotated reliably. For example, four piloting iterations were conducted before commencing the annotation of the PPI corpus. As a result, it was decided to remove MutationType from the list of originally proposed entity types as this information did not occur frequently enough in the piloting documents. The piloting process also helped to produce comprehensive annotation guidelines on all markables. During the piloting phase, the same documents were annotated by two or three annotators, IAA was computed for these documents, and annotation differences were analysed. The annotators discussed points of difficulty and disagreement with the NLP team and the annotation guidelines were clarified and extended wherever necessary.

At the end of the piloting phase a final set of markables was agreed by all parties and the main body of annotation commenced. During this phase weekly annotation meetings were held to discuss the latest IAA measurements and any other issues arising from the annotation, with all the annotators in attendance plus a representative from the NLP team. IAA was measured using a sample of documents randomly selected in advance for multiple annotation. The annotation was organised so that annotators were not aware when they were assigned a document that was being annotated by someone else as well. When new annotators joined the team they went through a training phase where they annotated several documents, comparing their annotations with those created by the existing team. This was done to ensure that they were following the guidelines correctly and were consistent with the other annotators.

For the annotation of the PPI corpus, an in-house annotation tool was developed using FilemakerPro, with data being stored in a relational database before being exported to XML for analysis by the NLP team. However, as this annotation tool did not scale well, a customised version of Callisto[13] was employed for the TE annotation project. Before the documents were presented to the annotators, they were tokenised and had sentence boundaries inserted by means of pre-processing steps implemented using the LT-XML2 and LT-TTT2 tools. The original spacing in the documents was preserved so that it could be recovered from the XML version simply by stripping off the word, sentence and paragraph elements.

All annotated documents were converted to an in-house XML format, for consumption by NLP applications. In the XML, all annotations are placed in standoff, with the normalisations included in the named entity annotation, and the properties and attributes included in the relation annotation. Listings 1, 2 and 3 show a sample of text, with its standoff entity and relation annotation. The standoff entity annotation uses word ids to refer to the start and end words of the entity, and the standoff relation annotation uses entity ids to refer to its entity pair. Note that the standoff markup for a document and its text are contained within the same file. An XML schema and format documentation will be provided with the corpus release.

```
<s><w id=”A33864”>Rrs1p</w>
 <w id=”A33870”>has</w> <w id=”A33874”>a</w>
 <w id=”A33876”>two</w><w id=”A33879”>-</w>
 <w id=”A33880”>hybrid</w>
 <w id=”A33887”>interaction</w>
 <w id=”A33899”>with</w> <w id=”A33904”>L5</w>
 <w id=”A33906”>.</w></s>
```

Listing 1: Extract from the text of an annotated document (note the original does not contain the line breaks)

[13] http://callisto.mitre.org/

Name	Entity type	Explanation	Count
ModificationBeforeEntity	Modification	Any modification applied before the interaction.	240
ModificationAfterEntity	Modification	Any modification resulting from the interaction.	1,198
DrugTreatmentEntity	DrugCompound	Any drug treatment applied to the interactors.	844
CellLineEntity	CellLine	The cell-line from which the interactor was drawn.	2,000
ExperimentalMethodEntity	ExperimentalMethod	The method used to detect the interactor.	1,197
MethodEntity	ExperimentalMethod	The method used to detect the interaction.	2,085
InteractionWordEntity	InteractionWord	The term which indicates the interaction.	11,386

Table 6: Attributes in the PPI corpus.

Name	Entity type	Explanation	Count
te_rel_ent-drug-compound	DrugCompound	Any drug compound applied.	1,549
te_rel_ent-exp-method1	ExperimentalMethod	The method used to detect the expression participants.	1,878
te_rel_ent-disease	DiseaseType	Any disease affecting the tissue.	332
te_rel_ent-dev-stage	DevelopmentalStage	The developmental stage of the tissue.	327
te_rel_ent-expr-word	ExpressionLevelWord	A term indicating the level of expression.	2,815

Table 7: Attributes in the TE corpus.

```
<ent id=”e933262” norm=”NP_014937” type=”Protein”
 species=”4932” sw=”A33864” ew=”A33864”>Rrs1p</ent>
<ent id=”e933263” norm=”” type=”ExperimentalMethod”
 sw=”A33876” ew=”A33880”>two−hybrid</ent>
<ent id=”e933264” norm=”” type=”InteractionWord”
 sw=”A33887” ew=”A33887”>interaction</ent>
<ent id=”e933265” norm=”NP_015194” conf=”100”
 type=”Protein” species=”4932” sw=”A33904”
 ew=”A33904”>L5</ent>
```

Listing 2: Example of standoff annotation of entities

```
<relation type=”ppi” id=”r903106” IsProven=”Proven”
 IsDirect=”Direct” IsPositive=”Positive”>
<argument ref=”e933262”></argument>
<argument ref=”e933265”></argument>
<attribute name=”MethodEntity” ref=”e933263”/>
<attribute name=”InteractionWordEntity”
 ref=”e933264”/>
</relation>
```

Listing 3: Example of standoff annotation of relations

Type	PPI		TE	
CellLine	81.6	(2,456)	—	
Complex	76.4	(2,243)	82.6	(886)
DevelopmentalStage	—		72.7	(357)
Disease	—		74.3	(435)
DrugCompound	76.4	(3,705)	84.9	(4,453)
ExperimentalMethod	74.0	(4,673)	76.7	(2,013)
Fragment	75.3	(3,985)	77.7	(1,179)
Fusion	78.5	(1,270)	73.9	(359)
GOMOP	—		50.2	(655)
Gene	—		77.7	(1,911)
Modification	87.6	(1,900)	—	
mRNAcDNA	—		78.1	(1,768)
Mutant	60.4	(1,008)	63.9	(310)
Protein	91.6	(32,799)	90.3	(16,329)
Tissue	—		84.1	(8,210)
All	84.9	(54,039)	83.8	(38,865)

Table 8: IAA for entities (in F_1) in each corpus. The total number of true positives is shown in brackets.

4.4 Inter-annotator Agreement

We IAA for each corpus and each markable using the multiply annotated documents. For each pair of annotations on the same document, IAA was calculated by scoring one annotator against another using precision, recall and F_1. For the PPI corpus, IAA was calculated on a total of 146 document pairs. IAA for TE corpus, having fewer triple annotations, was computed over a total of 92 document pairs. An overall corpus IAA was calculated by micro-averaging across all annotated document pairs.[14] Micro-averaging was chosen over macro-averaging, since we felt that the latter would give undue weight to documents with few or no markables. We used F_1 rather than Kappa (Cohen, 1960) to measure IAA since the latter requires comparison with a random baseline, which would not make sense for tasks such as named entity recognition and normalisation.

For named entities, IAA was calculated using precision, recall and F_1, defining two entities as equal if they had the same left and right boundaries, and the same type. The IAA figures for named entities listed in Table 8 show that annotation consistency is generally high, with important and frequently occurring entities scoring in the 80s or 90s. IAA is low for entity types which occur infrequently such as Mutant. It is particularly low for GOMOP, not only an infrequent entity but also an artificially constructed class designed to include cases of annotator uncertainty. The overall IAA is lower than that normally reported for MUC type entities, but fits with our observations that biomedical named entity annotation is more difficult.

The IAA for normalisations was only calculated when both annotators agreed on the entities. This means that the normalisation IAA only reflects agreement on normalisation annotation and is not affected by the level of agreement on the entity annotation. In addition, all entities marked as general were excluded from the IAA calculations (see Table 9). For Protein and mRNAcDNA types, only those entities that were normalised to RefSeq identifiers were included in the IAA calculations while for Gene and GOMOP entities, only those entities normalised to EntrezGene identifiers were included. The IAA was measured using F_1 where two normalisations were considered equal if both an-

[14] Micro-averaging means giving equal weight to each example, as opposed to macro-averaging which would give equal weight to each annotated document pair.

Type	PPI		TE	
DrugCompound	—		97.7	(215)
GOMOP	—		77.3	(214)
Gene	—		95.1	(1,463)
mRNAcDNA	—		88.0	(892)
Protein	88.4	(7,595)	90.0	(5,979)
Tissue	—		82.9	(6,776)
All	88.4	(7,595)	83.8	(15,785)

Table 9: IAA for normalisation (in F_1) in each corpus. The total number of true positives is shown in brackets.

Type	PPI		TE	
PPI	67.0	(2,729)	—	
TE	—		70.1	(2,078)
FRAG	84.6	(3,661)	84.0	(1,012)
All	76.1	(6,390)	74.1	(3,090)

Table 10: The IAA for relations (in F_1) in each corpus. The total number of true positives is shown in brackets. Note that FRAG relations are referred to as CHILD-PARENT in the TE corpus.

notators selected the same ID.

When calculating IAA for relations, only those relations for which both annotators agreed on the entities were included. Relation IAA was also measured using F_1, where relations are counted as equal if they connect exactly the same entity pair, and have the same type. The IAA for relations shown in Table 10 is overall lower than that for entities and normalisations, suggesting that this is a more difficult task. Since relations can span across clauses and even across sentences, the annotators need to perform a deeper analysis of the text than for entity annotations.

For properties, IAA was calculated for each name-value pair, again using precision, recall and F_1. In cases where the annotators had entered multiple relations of the same type between the same entities, these sets of equivalent relations were collapsed for the purpose of property and attribute IAA calculation. The collapsed relation was given the union of all the properties and attributes assigned to the relations in the set. This collapsing is an approximation of the annotator's intentions, but the number of occurrences of multiple equivalent relations is small so the collapsing should not have a significant effect on the IAA. The IAA for properties shown in Table 11 is generally very high, except for the IsProven-Unspecified category which was used infrequently by the annotators and suffers from being an "other" category.

For attributes, IAA was again measured using precision, recall and F_1. Two attributes were considered equivalent if they had the same type and connected the same relation and entity. Tables 12 and 13 show the IAA figures for attributes. These are quite low in some cases, and so are the total numbers of attributes assigned. Investigation of the IAA suggests that annotators often disagreed about whether to assign an attribute or not, but if they both assigned an attribute then they generally chose the same one. The entities used as attributes sometimes appeared at a distance from the relation in the text. Therefore, it is not surprising that annotators sometimes missed them, or assigned them inconsistently.

Name	Value	PPI		TE	
IsPositive	Positive	99.6	(2,553)	97.2	(1,807)
	Negative	90.1	(155)	88.9	(280)
IsDirect	Direct	86.8	(1,746)	—	
	NotDirect	61.4	(449)	—	
IsProven	Proven	87.8	(1,543)	92.8	(1,547)
	Referenced	88.6	(626)	75.3	(204)
	Unspecified	34.4	(448)	29.3	(38)
	All	87.2	(7.165)	91.2	(3,779)

Table 11: IAA for properties (in F_1) in each corpus. The total number of true positives is shown in brackets.

Name	IAA	
ModificationBeforeEntity	65.3	(31)
ModificationAfterEntity	86.7	(248)
DrugTreatmentEntity	45.4	(61)
CellLineEntity	64.0	(244)
ExperimentalMethodEntity	36.9	(94)
MethodEntity	55.4	(274)
All	59.6	(952)

Table 12: IAA of attributes (in F_1) in the PPI corpus. The total number of true positives is shown in brackets.

Name	IAA	
te_rel_ent-drug-compound	77.9	(229)
te_rel_ent-exp-method1	81.3	(261)
te_rel_ent-disease	64.0	(16)
te_rel_ent-dev-stage	57.8	(13)
All	77.2	(521)

Table 13: IAA of attributes (in F_1) in the TE corpus. The total number of true positives is shown in brackets.

5 Discussion and Conclusions

In terms of the amount of text annotated, the ITI TXM corpora are the result of one of the largest biomedical corpus annotation projects attempted to date. The two domains covered (protein-protein interactions and tissue expression) are both of crucial importance to biologists. Although there are several corpora already available with annotations of PPI, most of these only include protein annotation, and do not include the range of entities and normalisations available in the ITI TXM corpora. There are few available annotated corpora addressing tissue expression, and we are unaware of any large-scale efforts whose main focus is that domain.

Another interesting aspect of the ITI TXM corpora is the annotation of normalisations for multiple types of entity mentions, and for multiple species. This annotation was motivated by the role of the NLP system, as an assistant to curators, as it was suspected that mapping proteins, genes and other terms to standard databases occupied a significant proportion of curators' time. The annotation of multi-species normalisations was difficult in situations where it was unclear which species was being referred to for a given named entity mention. These issues were resolved by deriving a series of annotation guidelines, as detailed in Section 4.2. The annotation guidelines were also reasonably successful in ensuring annotator consistency, as evidenced by the normalisation IAA provided in Section 4.4.

During the annotation we found that the interaction between the NLP team and the biologists was essential at all stages. In the design phase, the biologists, as the domain experts, provided insight into what information should be annotated. At the same time, the NLP team were able to explain to the biologists what their technology is capable of. However, although both parties have an insight into what can be reliably annotated, the only sure way to determine this is empirically through extensive piloting. The piloting phase not only provided experimental data on annotation agreement and timing, but also helped the NLP team and the biologists to improve their shared understanding of the annotation process and its difficulties. During the main annotation phase, it was helpful to have regular contact between the NLP and the annotation teams in order to ensure that doubts and difficulties were noted, discussed and resolved as quickly as possible. The NLP team analysed the data as it was produced by the annotators and drew their attention to any recurring sources of disagreement.

We believe that measuring IAA is a crucial part of any corpus annotation effort. It provides a check that the annotators are producing a reliable and consistent corpus. It also gives a measure of how difficult the task is and suggests how well an automated system can be expected to perform. We took steps to ensure that the IAA itself was reliable, by instructing annotators not to discuss papers whilst annotating them. We also did not inform annotators in advance whether they were working on a paper that was also being annotated by another person. The IAA measurements for the final set of markables shows that some proved difficult to annotate reliably, for example the GOMOP entity and some of the attributes. Annotating them was problematic in the piloting phase, and whilst we attempted to tighten up the guidelines, it was not sufficient to boost their IAA.

We hope that the two ITI TXM corpora, consisting of over 200 papers each, and with multiple types of semantic annotation, will provide a useful resource for the biomedical text-mining community when released to the academic research community later this year.

6 Acknowledgements

The ITI TXM corpora were created as part of an ITI Life Sciences Scotland (http://www.itilifesciences.com) research programme with Cognia EU and the University of Edinburgh. We would like to thank the annotation teams led by Elizabeth Fairley and Lynn Morrice as well as Cognia EU's software development team. Thanks are also due to Malvina Nissim, Kirsten Lillie and Henk Harkema for their help with devising the annotation guidelines and deciding on sets of markables.

7 References

Beatrice Alex, Malvina Nissim, and Claire Grover. 2006. The impact of annotation on the performance of protein tagging in biomedical text. In *Proceedings of LREC*.

Razvan Bunescu, Ruifang Ge, Rohit J. Kate, Edward M. Marcotte, Raymond J. Mooney, Arun K. Ramani, and Yuk W. Wong. 2005. Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine*, 33(2):139–155.

Jean Carletta, David McKelvie, Amy Isard, Andreas Mengel, Marion Klein, and Morton Baun Møller. 2005. A generic approach to software support for linguistic annotation using XML. In Geoffrey Sampson and Diana McCarthy , editors, *Readings in Corpus Linguistics*. Continuum International.

Kevin B. Cohen, Lynne Fox, Philip V. Ogren, and Lawrence Hunter. 2005. Corpus design for biomedical natural language processing. In *Proceedings of ISMB*.

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

FetchProt, 2005. *The FetchProt Corpus: documentation and annotation guidelines*. Available online at: http://fetchprot.sics.se.

Claire Grover, Michael Matthews, and Richard Tobin. 2006. Tools to address the interdependence between tokenisation and standoff annotation. In *Proceedings of NLPXML*.

Martin Krallinger, Rainer Malik, and Alfonso Valencia. 2006. Text mining and protein annotations: the construction and use of protein description sentences. *Genome Inform*, 17(2):121–130.

Seth Kulick, Ann Bies, Mark Liberman, Mark Mandel, Ryan Mcdonald, Martha Palmer, Andrew Schein, Lyle Ungar, Scott Winters, and Pete White. 2004. Integrated annotation for biomedical information extraction. In *Proceedings of the BioLINK*.

Zhiyong Lu, Michael Bada, Philip V. Ogren, K. Bretonnel Cohen, and Lawrence Hunter. 2006. Improving biomedical corpus annotation guidelines. In *Proceedings of the Joint BioLINK and 9th Bio-Ontologies Meeting*.

Inderjeet Mani, Zhangzhi Hu, Seok Bae Jang, Ken Samuel, Matthew Krause, Jon Phillips, and Cathy H. Wu. 2005. Protein name tagging guidelines: lessons learned. *Comparative and Functional Genomics*, 6(1-2):72–76.

Tara McIntosh and James R. Curran. 2007. Challenges for extracting biomedical knowledge from full text. In *Proceedings of BioNLP*.

Claire Nedellec. 2005. Learning language in logic - genic interaction extraction challenge. In *Proceedings of the ICML Workshop on Learning Language in Logic*.

Tomoko Ohta, Yuka Tateisi, and Jin-Dong Kim. 2002. GENIA corpus: an annotated research abstract corpus in molecular biology domain. In *Proceedings of HLT*.

Scott Piao, Ekaterina Buyko, Yoshimasa Tsuruoka, Katrin Tomanek, Jin-Dong Kim, John McNaught, Udo Hahn, and Sophia Ananiadou. 2007. BootStrep annotation scheme - encoding information for text mining. Proceedings of the 4th Corpus Linguistics Conference.

Sampo Pyysalo, Filip Ginter, Juho Heimonen, Jari Björne, Jorma Boberg, Jouni Järvinen, and Tapio Salakoski. 2007. BioInfer: A corpus for information extraction in the biomedical domain. *BMC Bioinformatics*, 8(1).

Parantu K. Shah, Carolina Perez-Iratxeta, Peer Bork, and Miguel A. Andrade. 2003. Information extraction from full text scientific articles: where are the keywords? *BMC Bioinformatics*, 4(20).

Hagit Shatkay and Ronen Feldman. 2003. Mining the biomedical literature in the genomic era: an overview. *Journal of Computational Biology*, 10(6):821–855.

Lorraine Tanabe, Natalie Xie, Lynne H. Thom, Wayne Matten, and W. John Wilbur. 2005. GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, 6 Suppl 1.

GENIA Treebank, 2005. *GENIA Treebank Beta Version*. Available online at: http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/topics/Corpus/GTB.html.

John W. Wilbur, Andrey Rzhetsky, and Hagit Shatkay. 2006. New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinformatics*, 7(1).

Semantic Annotation of Clinical Text: The CLEF Corpus

Angus Roberts, Robert Gaizauskas, Mark Hepple, George Demetriou, Yikun Guo, Andrea Setzer, Ian Roberts
Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello, Sheffield, UK S1 4DP
E-mail: initial.surname@dcs.shef.ac.uk

Abstract

A significant amount of important information in Electronic Health Records (EHRs) is often found only in the unstructured part of patient narratives, making it difficult to process and utilize for tasks such as evidence-based health care or clinical research. In this paper we describe the work carried out in the CLEF project for the semantic annotation of a corpus to assist in the development and evaluation of an Information Extraction (IE) system as part of a larger framework for the capture, integration and presentation of clinical information. The CLEF corpus consists of both structured records and free text documents from the Royal Marsden Hospital pertaining to deceased cancer patients. The free text documents are of three types: clinical narratives, radiology reports and histopathology reports. A subset of the corpus has been selected for semantic annotation and two annotation schemes have been created and used to annotate: (i) a set of clinical entities and the relations between them, and (ii) a set of annotations for time expressions and their temporal relations with the clinical entities in the text. The paper describes the make-up of the annotated corpus, the semantic annotation schemes used to annotate it, details of the annotation process and of inter-annotator agreement studies, and how the annotated corpus is being used for developing supervised machine learning models for IE tasks.

1. Introduction

Although large parts of the patient electronic health care record exist as structured data, a significant proportion exists as unstructured free texts. This is not just the case for legacy records. Much of pathology and imaging reporting is recorded as free text, and a major component of any UK medical record consists of letters written from the secondary to the primary care physician (GP). These documents contain information of value for day-to-day patient care and of potential use in research. For example, narratives record why drugs were given, why they were stopped, the results of physical examination, and problems that were considered important when discussing patient care, but not important when coding the record for audit. Clinical researchers could be assisted in hypothesis formation (for subsequent verification in clinical trials) if they could get answers aggregated across all NHS patient records to questions such as:

> *How many patients with stage 2 adenocarcinoma who were treated with tamoxifen were symptom-free after 5 years?*

Doctors could also benefit for treating individual patients if they could get concise summaries of patients' clinical histories or if they had access to histories of similar patients elsewhere.

CLEF (Rector et al. 2003) uses IE technology to make information available for integration with the structured record, and thus to make it available for clinical care and research (Harkema et al. 2005). Although some IE research has focused on unsupervised methods of developing systems, as in the earlier work of Riloff (1996), most practical IE still needs data that has been manually annotated with events, entities and relationships. This data serves three purposes. Firstly, an analysis of human annotated data focuses and clarifies requirements. Secondly, it provides a gold standard against which to assess results. Thirdly, it provides data for system development: extraction rules may be created either automatically or by hand, and statistical models of the text may be built by machine learning algorithms.

This paper reports on the construction of a gold standard corpus for the CLEF project, in which clinical documents are annotated both with multiple entities and their relationships. To the best of our knowledge, no one has explored the problem of producing a corpus annotated for clinical IE to the depth and to the extent reported here. Our annotation exercise uses a large corpus, covers multiple text types, and involves over 20 annotators. We examine two issues of pertinence to the annotation of clinical documents: the use of domain knowledge; and the applicability of annotation to different sub-genres of text. Results are encouraging, and suggest that a rich corpus to support IE in the medical domain can be created. An earlier description of the CLEF corpus was reported in (Roberts et al. 2007). The current paper provides more details, including details of the temporal annotation (not reported at all earlier), figures on the distribution of entity and relation types across the corpus, and inter annotator agreement scores for the completed corpus.

The next section of this paper summarises the literature about annotated biomedical corpora. The following section describes the design of the CLEF corpus, describing the selection of documents for gold standard semantic annotation and the entities and relationships with which the gold standard is annotated. Next the annotation methodology is described, including a discussion of the development of annotation guidelines and an assessment of the consistency of human annotations. The following sections present inter annotator agreement scores for the finished corpus, and figures on the distribution of entity and relation types by document type across the corpus. Finally we mention on-going use of the corpus for training and evaluation of our supervised machine learning IE system.

2. Annotated Corpora for Biomedical Research

Semantically annotated corpora are becoming increasingly common within biomedical information extraction research, with annotation levels gradually expanding over the years. For example, the GENIA corpus of Medline abstracts has been annotated with information about biological entities (Kim et al. 2003) with annotations about biological events added to (part of) it at a later stage (Kim et al. 2008). Other semantically annotated corpora developed for the purpose of providing training and evaluation material for IE systems include:

- The PennBioIE corpus of ~2300 Medline abstracts, in the domains of molecular genetics of oncology and inhibition of enzymes of the CYP450 class annotated for biomedical entity types and parts-of-speech, some of which have also been annotated for Penn Treebank style syntactic structure (Mandel, 2006);
- The Yapex corpus of 200 Medline abstracts annotated for protein names (Franzén et al. 2002);
- Those developed within the BioText project for disease-treatment relation classification (Rosario and Hearst, 2004) and protein-protein interaction classification (Rosario and Hearst, 2005).

In addition corpora have been available in order to provide data sets for research competitions such as:

- Biocreative (the GENETAG corpus containing 15,000 sentences with gene/protein names annotated – Tanabe et al 2005)
- the TREC Genomics Track, which ran from 2003-2007 and for which a variety of datasets and tasks were developed (http://ir.ohsu.edu/genomics/).
- the LLL05 challenge task, which supplied training and test data for the task of identifying protein/gene interactions in sentences from Medline abstracts (Nédellec, 2005).

All of the above corpora consist of texts drawn from the research literature, in most cases from the biology research literature. This is due at least in part to the difficulty of getting access to clinical text for research purposes. To our knowledge the only other work in the area of corpus annotation for clinical information retrieval and extraction is:

- The corpus prepared and released for the Computational Medicine Challenge (Pestian et al 2007). This corpus consists of 1954 (978 training, 976 test) radiology reports annotated with ICD-9-CM codes, the challenge being the text classification challenge of automatically coding the unseen test data.
- The ImageCLEFmed 2005 and 2006 image test collections which consist of ~50,000 images with associated textual annotations (case descriptions, imaging reports) and in some cases metadata (e.g. DICOM labels), together with query topics and relevance judgements (Hersh et al 2006; Müller et all 2007). While intended to support medical image retrieval research, the textual component of this resource could have purely language processing applications.
- Ogren et al.'s (2006) work on annotating disorders within clinic notes; and
- The I2B2 challenges, which have so far provided training and evaluation data for de-identification of discharge summaries and for the identification of smoking status from discharge summaries (challenge 1); and for identification of obesity and co-morbidities from discharge summaries annotated at the document level (https://www.i2b2.org/NLP/).

What differentiates CLEF from the annotation exercises mentioned above is that (1) it is the only corpus annotated with information about clinical entities and their relations as well as with temporal information about the clinical entities and time expressions occurred in patient narratives and (2) it is the only corpus to contain clinic notes, radiology reports and histopathology reports together with associated structured data.

3. Design of the CLEF Corpus

Our development corpus comes from CLEF's main clinical partner, the Royal Marsden Hospital, a large specialist oncology centre. The entire corpus consists of both the structured records and free text documents from 20234 deceased patients. The free text documents consist of three types: clinical narratives (with sub-types as shown in Table 1); histopathology reports; and imaging reports. Patient confidentiality is ensured through a variety of technical and organisational measures, including automatic pseudonymisation and manual inspection. Approval to use this corpus for research purposes within CLEF was sought and obtained from the Thames Valley Multi-centre Research Ethics Committee (MREC).

3.1 Gold Standard Document Sampling

Given the expense of human annotation, the gold standard portion of the corpus has to be a relatively small subset of the whole corpus of 565000 documents. In order to avoid events that are either rare or outside of the main project requirements, it is restricted by diagnosis, and only considers documents from those patients with a primary diagnosis code in one of the top level sub-categories of ICD-10 Chapter II (neoplasms). In addition, it only contains those sub-categories that cover more than 5% of narratives and reports. The gold standard corpus consists of two portions, selected for slightly different purposes.

3.1.1 Whole patient records

Two applications in CLEF involve aggregating data across a single patient record. The CLEF chronicle builds a chronological model for a patient, integrating events from both the structured and unstructured record (Rogers et al 2006). CLEF report generation creates aggregated and natural language reports from the chronicle (Hallet et al 2006). These two applications require whole patient records for development and testing. Two whole patient records were selected for this portion of the corpus, from two of the major diagnostic categories, to give median numbers of documents, and a mix of document types and lengths. Each record consists of nine narratives, one radiology report and seven histopathology reports, plus associated structured data.

3.1.2 Stratified random sample

The major portion of the gold standard serves as development and evaluation material for IE. In order to ensure even training and fair evaluation across the entire corpus, the sampling of this portion is randomised and stratified, so that it reflects the population distribution along various axes. Table 1 shows the proportions of clinical narratives along two of these axes. The random sample consists of 50 each of clinical narratives, histopathology reports, and imaging reports.

Narrative subtype	% of standard
To GP	49
Discharge	17
Case note	15
Other letter	7
To consultant	6
To referrer	4
To patient	3

Neoplasm	% of standard
Digestive	26
Breast	23
Haematopoetic	18
Respiratory etc	12
Female genital	12
Male genital	8

Table 1: % of narratives in random sample

3.2 Annotation Schema: Clinical Information

The CLEF gold standard is a semantically annotated corpus. We are interested in extracting the main semantic entities from text. By *entity*, we mean some real-world concept referred to in the text such as the drugs that are mentioned, the tests that were carried out etc. We are also interested in extracting the *relationships* between entities: the condition indicated by a drug, the result of an investigation etc.

Annotation is anchored in the text. Annotators mark spans of text with a type: drug, locus and so on. Annotators may also mark words that modify spans (such as negation), and mark relationships as links between spans. Two or more spans may refer to the same thing in the real world, in which case they *co-refer*. Co-referring CLEF entities are linked by the annotators.

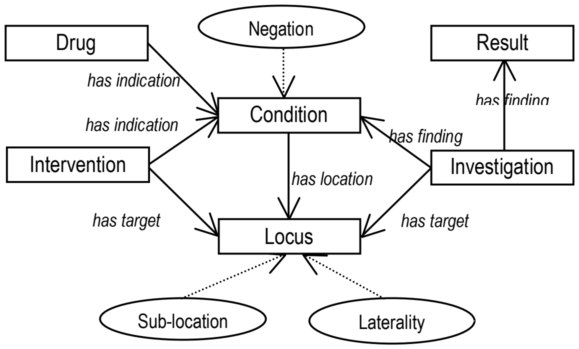


Figure 1: CLEF annotation schema. Rectangles: entities; ovals: modifiers; solid lines: relationships.

The types of annotation are described in a schema, shown in Figure 1. The CLEF entities and relations are also listed in Tables 2 and 3, along with descriptions and examples.

The schema has been based on a set of requirements developed between clinicians and computational linguists in CLEF. The schema types are mapped to types in the UMLS semantic network, which enables us to utilize UMLS vocabularies in entity recognition. For the purposes of annotation, the schema is modeled as a Protégé-Frames ontology (Gennari et al. 2003). Annotation is carried out using an adapted version of the Knowtator plugin for Protégé (Ogren 2006). This was chosen for its handling of relationships, after evaluating several such tools.

3.3 Annotation Schema: Temporal Information

Information from structured data and clinical narratives is integrated to build a *patient chronicle*, i.e. a coherent overview of the significant events in the patients' medical history, such as their condition, diagnosis and treatment over the period of care. This process involves extracting temporal information about events from the narratives, and using this and other information to map the events extracted from the narratives onto their corresponding, time-stamped, events in the structured data wherever possible. The aim of the gold standard is to provide the temporal links (called CTlinks for CLEF Temporal link) between TLCs (Temporally Located CLEF entities, which comprise CLEF investigations, interventions and conditions) and temporal expressions. Temporal expressions include dates and times (both absolute and relative), as well as durations, as specified in the TimeML (2004) TIMEX3 standard. CTlinks types include, for example, before, after, overlap, includes (for a full list see Table 9). Our scheme requires annotation of only those temporal relations holding between TLCs and the date of the letter (Task A), and between TLCs and temporal expressions appearing in the same sentence (Task B). These tasks are similar to, but not identical with, those addressed by the TempEval challenge within SemEval 2007 (Verhagen et al. 2007).

4. Annotation Methodology

The annotation methodology follows established natural language processing standards (Boisen et al. 2000). Annotators work to agreed guidelines; documents are annotated by at least two annotators; documents are only used where agreement passes a threshold; differences are resolved by a third experienced annotator. These points are discussed further below.

Entity type	Description	Example
Condition	Symptom, diagnosis, complication, conditions, problems, functions and processes, injury	*This patient has had a lymph node biopsy which shows melanoma in his right groin. It is clearly secondaries from the melanoma on his right second toe.*
Intervention	Action performed by doctor or other clinician targeted at a patient, **Locus**, or **Condition** with the objective of changing (the properties) of, or treating, a **Condition**.	*Although his PET scan is normal he does need a groin dissection* *We agreed to treat with DTIC, and then consider radiotherapy*.
Investigation	Interaction between doctor and patient or **Locus** aimed at measuring or studying, but not changing, some aspect of a **Condition**. **Investigations** have findings or interpretations, whereas **Interventions** usually do not.	*This patient has had a lymph node biopsy … Although his PET scan is normal he does need a groin dissection. We will perform a CT scan to look at the left pelvic side wall …*
Result	The numeric or qualitative finding of an **Investigation**, excluding **Condition**	*Although his PET scan is normal…* Other examples include the numeric values of tests, such as "80mg".
Drug or device	Usually a drug. Occasionally, medical devices such as suture material and drains will also be mentioned in texts.	*This (pain) was initially relieved by co-codamol*
Locus	Anatomical structure or location, body substance, or physiologic function, typically the locus of a **Condition**.	*This patient has had a lymph node biopsy which shows melanoma in his right groin … It is clearly secondaries from the melanoma on his right second toe. Although his PET scan is normal he does need a groin dissection. We will perform a CT scan to look at the left pelvic side wall*

Table 2: CLEF Entities

Relation type	1st arg type	2nd arg type	Description	Example
has_target	**Investigation, Intervention**	**Locus**	Relates an intervention or an investigation to the bodily locus at which it is targetted.	*This patient has had a [arg2 lymph node] [arg1 biopsy]* *… he does need a [arg2 groin] [arg1 dissection]*
has_finding	**Investigation**	**Condition, Result**	Relates a condition to an investigation that demonstrated its presence, or a result to the investigation that produced that result.	*This patient has had a lymph node [arg1 biopsy] which shows [arg2 melanoma]* *Although his [arg1 PET] scan is [arg2 normal]*
has_indication	**Drug or device, Intervention, Investigation**	**Condition**	Relates a condition to a drug, intervention, or investigation that is targetted at that condition	*Her facial [arg2 pain] was initially relieved by [arg1 co-codamol]*
has_location	**Condition**	**Locus**	Relationship between a condition and a locus: describes the bodily location of a specific condition. May also describe the location of malignant disease in lymph nodes, relating an involvement to a locus.	*… a biopsy which shows [arg1 melanoma] in his right [arg2 groin]* *It is clearly secondaries from the [arg1 melanoma] on his right [arg2 second toe]* *Her[arg2 facial] [arg1 pain] was initially relieved by co-codamol*
Modifies	**Negation signal**	**Condition**	Relates a condition to its negation or uncertainty about it	*There was [arg1 no evidence] of extra pelvic [arg2 secondaries]*
Modifies	**Laterality signal**	**Locus, Intervention**	Relates a bodily locus or intervention to its sidedness: *right*, *left*, *bilateral*.	*… on his [arg1 right] [arg2 second toe]* [arg1 *right*] [arg2 *thoracotomy*]
Modifies	**Sub-location signal**	**Locus**	Relates a bodily locus to other information about the location: *upper*, *lower*, *extra*, etc.	*[arg1 extra] [arg2 pelvic]*

Table 3: CLEF Relations

4.1 Annotation Guidelines

Consistency is critical to the quality of a gold standard. It is important that all documents are annotated to the same standard. Questions regularly arise when annotating. For example, should multi-word expressions be split? Should “myocardial infarction” be annotated as a condition, or as a condition and a locus? To ensure consistency, a set of guidelines is provided to annotators. These describe in detail what should and should not be annotated; how to decide if two entities are related; how to deal with co-reference; and a number of special cases. The guidelines also provide a sequence of steps, a recipe, which annotators should follow when working on a document. This recipe is designed to minimise errors of omission. The guidelines themselves were developed through a rigorous, iterative process, which is described below.

4.2 Double Annotation

A singly annotated document can reflect many problems: the idiosyncrasies of an individual annotator; one-off errors made by a single annotator; annotators who consistently under-perform. There are many alternative annotation schemes designed to overcome this, all of which involve more annotator time. Double annotation is a widely used alternative, in which each document is independently annotated by two annotators, and the sets of annotations compared for agreement.

4.3 Agreement Metrics

We measure agreement between double annotated documents using *inter annotator agreement* (IAA, shown below).

IAA = matches / (matches + non-matches)

We report IAA as a percentage. Overall figures are macro-averaged across all entity or relationship types. Entity IAA may be either “relaxed” or “strict”. In relaxed IAA, partial matches, i.e. overlaps, are counted as a half match. In strict IAA, partial matches do not count to the score. Together, these show how much disagreement is down to annotators finding similar entities, but differing in the exact spans of text marked. We used both scores in development, but provide a set of final strict IAAs for the corpus. Results given below explicitly state the score being used.

Two variations of relationship IAA were used. First, all relationships found were scored. This has the drawback that an annotator who failed to find a relationship because they had not found one or both the entities would be penalized. To overcome this, a Corrected IAA (referred to as CIAA) was calculated, including only those relationships where both annotators had found the two entities involved. This allows us to isolate, to some extent, relationship scoring from entity scoring.

4.4 Difference Resolution

Double annotation can be used to improve the quality of annotation, and therefore the quality of statistical models trained on those annotations. This is achieved by combining double annotations to give a set closer to the "truth" (although it is generally accepted as impossible to define an "absolute truth" gold standard in an annotation task with the complexity of CLEF's). The resolution process is carried out by a third experienced annotator. All agreements from the original annotators are accepted into a consensus set, and the third annotator adjudicates on differences, according to a set of strict guidelines. In this way, annotations remain at least double annotated.

4.5 Developing the Guidelines

The guidelines were developed and refined using an iterative process, designed to ensure their consistency. This is shown in Figure 2. Two qualified clinicians annotated different sets of documents in 5 iterations (covering 31 documents in total). The relaxed IAA and CIAA for these iterations are shown in Table 4. As can be seen, entity relaxed IAA remains consistently high after the 5 iterations, after which very few amendments were required on the guidelines. Relation CIAA does not appear so stable on iteration 5. Difference analysis showed this to be due to a single, simple type of disagreement across a limited number of sentences in one document. Scoring without this document gave a 73% CIAA.

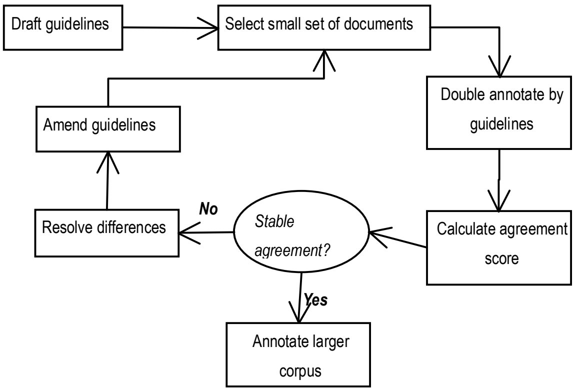


Figure 2: Iterative development of guidelines

		Debug iteration				
		1	2	3	4	5
Entities	Matches	244	244	308	462	276
	Partial match	2	6	22	6	1
	Non-matches	45	32	93	51	22
	Relaxed IAA	**84**	**87**	**74**	**89**	**92**
Relationships	Matches	170	78	116	412	170
	Partial match	3	5	14	6	1
	Non-matches	31	60	89	131	103
	Corrected IAA	**84**	**56**	**56**	**75**	**62**

Table 4: Relaxed IAA and CIAA (%) for each development iteration.

4.6 Annotator Expertise

In order to examine how easily the guidelines could be applied by other annotators with varying levels of expertise, we also gave a batch of documents to our development annotators, another clinician, a biologist with some linguistics background, and a computational linguist. Each was given very limited training. The resultant annotations were compared with each other, and with a consensus set created from the two development annotators. The relaxed IAA matrix for this group is shown in Table 5. This small experiment shows that even with very limited training, agreement scores that approach acceptability are achievable. A difference analysis suggested that the computational linguist was finding more pronominal co-references and verbally signaled relations than the clinicians, but that unsurprisingly, the clinicians found more relations requiring domain knowledge to resolve. A combination of both linguistic and medical knowledge appears to be best.

This difference reflects a major issue in the development of the guidelines: the extent to which annotators should apply domain specific knowledge to their analysis. Much of clinical text can be understood, even if laboriously and simplistically, by a non-clinician armed with a medical dictionary. The basic meaning is exposed by the linguistic constructs of the text. Some relationships between entities in the text, however, require deeper understanding. For example, the condition for which a particular drug was given may be unclear to the non-clinician. In writing the guidelines, we decided that such relationships should be annotated, although this requirement is not easy to formulate as specific rules.

D2	77				
C	67	68			
B	76	80	69		
L	67	73	60	69	
Consensus	**85**	**89**	**68**	**78**	**73**
	D1	D2	C	B	L

Table 5: Relaxed IAA (%) for entities. D1 and D2: development annotators; C: clinician; B: biologist with linguistics background; L: computational linguist

4.7 Annotation: Training and Consistency

In total, around 25 annotators were involved in guideline development and annotation. They included practicing clinicians, medical informaticians, and final year medical students. They were each given an initial 2.5 hours training session.

After the initial training session, annotators were given two training batches to annotate, which comprised documents originally used in the debugging exercise, and for which consensus annotations had been created. Relaxed IAA and CIAA were computed between annotators, and against the consensus set. These figures allowed us to identify and offer remedial training to under-performing annotators and to refine the guidelines further.

4.8 Annotation of Temporal Information

This work is still at a preliminary stage. To date ten patient letters (narrative data) for a number of patients have been annotated in accordance with the scheme described in Section 3.3 above, which we still view as under development. A second annotator is currently re-annotating as part of the guideline development phase (see Figure 2). Temporal annotation is done through a combination of manual and automatic methods. TLCs were imported from the part of the corpus already annotated with clinical entities. Temporal expressions were annotated and normalized to ISO dates by the GUTime tagger (Mani and Wilson 2000), developed at Georgetown University, which annotates in accordance with the TIMEX3 standard and also recognizes a variety of temporal modifiers and European date formats. After these automatic steps, we manually annotate the temporal relations holding between TLCs and the date of the letter (Task A), and between TLCs and temporal expressions appearing in the same sentence (Task B**).**

5. Inter annotator agreement

We have calculated IAA for the double annotations across the complete stratified random portions of the gold standard, for each document type. Table 6 shows the strict IAA for entities, and Table 7 shows both the IAA and CIAA for relationships.

Entity	**Narratives**	**Histopath.**	**Radiology**
Condition	81	67	77
Drug or device	84	59	32
Intervention	64	57	43
Investigation	77	56	70
Locus	78	71	75
Result	69	29	48
Laterality	95	88	91
Negation	67	71	65
Sub-location	63	29	36
Overall	**77**	**62**	**69**

Table 6: Strict IAA (%) for entities across the stratified random corpus

Note that the final gold standard consists of a consensus of the double annotation, created by a third annotator. Systems trained and evaluated with the gold standard use this consensus. The IAAs given do not therefore provide an upper bound on system performance, but an indication of how hard a recognition task is. Table 7 illustrates that relation annotation is highly dependent on entity annotation: CIAA, corrected for entity recognition, is significantly higher than uncorrected IAA.

	Narratives		**Histopath.**		**Radiology**	
Relation	**IAA**	**CIAA**	**IAA**	**CIAA**	**IAA**	**CIAA**
has_finding	48	76	26	69	33	55
has_indication	35	51	15	30	14	22
has_location	59	80	44	70	45	77
has_target	45	64	20	47	67	81
laterality_mod	70	93	70	89	55	80
negation_mod	63	90	67	100	51	94
sub_loc_mod	52	98	29	100	32	93
Overall	**52**	**75**	**36**	**72**	**43**	**76**

Table 7: IAA and corrected IAA (%) for relationships across the stratified random corpus

6. Distribution of semantic annotations

The distribution of annotations for CLEF entities and relations in the stratified random portion of the corpus (50 documents of each type) is shown in Table 8.

CLEF stratified random corpus				
Entity	Narratives	Histopathology	Radiology	Total
Condition	429	357	270	1056
Drug or device	172	12	13	197
Intervention	191	53	10	254
Investigation	220	145	66	431
Laterality	76	14	85	175
Locus	284	357	373	1014
Negation	55	50	53	158
Result	125	96	71	292
Sub-location	49	77	125	251
Relation				
has_finding	233	263	156	652
has_indication	168	47	12	227
has_location	205	270	268	743
has_target	95	86	51	232
laterality_mod	73	14	82	169
negation_mod	67	54	59	180
sub_loc_mod	43	79	125	247

Table 8: Distribution of annotations by document type for entities and relations (clinical IE).

The distribution of annotations for the different subtypes of CTLinks, TLCs and time expressions for the ten development documents annotated so far are shown in Tables 9 and 10. Note that some TLCs are marked as hypothetical. For example in "no palliative chemotherapy or radiotherapy would be appropriate" the terms chemotherapy and radiotherapy are marked as TLCs but clearly have no "occurrence" that can be located in time and hence will not participate in any CTLinks.

CTLink	Task A	Task B
After	5	18
Ended_by	3	0
Begun_by	4	0
Overlap	7	26
Before	5	135
None	4	8
Is_included	31	67
Unknown	6	14
Includes	13	137
Total	78	405

Table 9: Distribution of CTLinks by type for tasks A & B.

TLCs	Not hypothetical	243
	hypothetical	16
	Total	259
Time Expression	Duration	3
	DATE	52
	Total	55

Table 10: Distribution of TLCs and temporal expressions.

7. Using the Corpus

The gold standard corpus is used as input to train an IE system based on SVM classifiers for recognizing both entities and relations. Preliminary results, with models evaluated on a narrative corpus comprising the combined stratified random and whole patient portion, achieve average F-measure 71% for entity extraction over 5 entity types (Roberts et al. 2008). Preliminary results for relation extraction trained with the same corpus, achieve an average F-measure of 70% over 7 relation types, where gold standard entities are provided as input.

8. Conclusion

We have described the CLEF corpus: a semantically annotated corpus designed to support the training and evaluation of information extraction systems developed to extract information of clinical significance from free text clinic notes, radiology reports and histopathology reports. We have described the design of the annotated corpus, including the number of texts it contains, the principles by which they were selected from a large body of unannotated texts and the annotation schema according to which clinical and temporal entities and relations of significance have been annotated in the texts. We also described the annotation process we have undertaken with a view to ensuring, as far as is possible given constraints of time and money, the quality and consistency of the annotation, and we have reported results of inter-annotator agreement, which show that promising levels of inter-annotator agreement can be achieved. We have examined the applicability of annotation guidelines to several clinical text types, and our results suggest that guidelines developed for one type may be fruitfully applied to others. We have also reported the distribution of entity and relation types, both clinical and temporal, across the corpus, giving a sense of how well represented each entity and relation type is in the corpus.

The annotated CLEF corpus is the richest resource of semantically marked up clinical text yet created. Our work has faced several challenges, such as achieving consistent annotation, particularly of relations, across annotators and co-ordinating the work of many annotators at several sites. We do not as yet have persmission to release these materials to the wider language processing community for research purposes. However, we are currently preparating an application requesting this release, to be submitted shortly to the appropriate UK Multi-centre Research Ethics Committee. We are optimistic of success.

9. Acknowledgements

CLEF is funded by the UK Medical Research Council. We would like to thank the Royal Marsden Hospital for providing the corpus; our annotators; and members of CLEF who have helped with clinical expertise and logistics.

10. References

Boisen S, Crystal MR, Schwartz R, Stone R, Weischedel R. (2000). Annotating resources for information extraction. *Proc Language Resources and Evaluation, pp.* 1211--1214.

Franzén K., Eriksson G., Olsson F., Per Lidén L. A. and Cöster J. (2002). Protein names and how to find them. In

International Journal of Medical Informatics special issue on Natural Language Processing in Biomedical Applications.

Gennari J.H., Musen M.A., Fergerson R.W. et al. (2003). The evolution of Protégé: an environment for knowledge-based systems development. *Int J Hum Comput Stud.*, 58, pp. 89--123.

Hallet C., Power R., Scott D. (2006). Summarisation and visualization of e-health data repositories. *Proc 5th UK e-Science All Hands Meeting.*

Harkema H., Roberts I., Gaizauskas R., and Hepple M. (2005). Information extraction from clinical records. *Proc 4th UK e-Science All Hands Meeting.*

Hersh W.R., Müller H., Jensen J., Yang J., Gorman P., Ruch P. (2006). Advancing biomedical image retrieval: development and analysis of a test collection. *J Am Med Inform Assoc.*, 13, pp. 488--496.

Kim J-D, Ohta T, Tateisi Y, Tsujii J. (2003). GENIA corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics,* 19(1), pp. 180--182.

Kim J-D., Tateisi Y, Tsujii J. (2008). Corpus annotation for mining biomedical events from literature. *BMC Bioinformatics,* 9:10.

Mandel M. (2006). Integrated Annotation of Biomedical Text: Creating the PennBioIE corpus. *Text Mining, Ontologies and Natural Language Processing in Biomedicine*, Manchester, UK.

Mani I. and Wilson G. (2000). Processing of News. *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL2000)*, pp. 69--76.

Müller H., Deselaers T., Deserno T. et al (2007). Overview of the ImageCLEFmed 2006 Medical Retrieval and Medical Annotation Tasks. *Evaluation of Multilingual and Multi-modal Information Retrieval.* LNCS 4730/2007, Springer, pp. 595-608.

Nédellec, C. (2005). Learning Language in Logic - Genic Interaction Extraction Challenge. *Proceedings of the ICML05 Workshop on Learning Language in Logic*, Bonn. pp. 31-37.

Ogren P.V., Savova G.K., Buntrock J.D., Chute C.G. (2006). Building and evaluating annotated corpora for medical NLP systems. *Proc AMIA Annual Symposium.*

Ogren, P.V. (2006). Knowtator: a Protégé plug-in for annotated corpus construction. *Proc Human Language Technology*, pp. 273--275.

Pestian JP, Brew C, Matykiewicz PM, Hovermale DJ, Johnson N, Cohen KB, Duch W. (2007). A shared task involving multi-label classification of clinical free text. Proc. ACL BioNLP 2007, Prague.

Rector A., Rogers J., Taweel A. et al. (2003). CLEF: joining up healthcare with clinical and post-genomic research. *Proc 2nd UK e-Science All Hands Meeting.* pp. 264--267.

Riloff E. (1996), Automatically generating extraction patterns from untagged text. *Proc 13th Nat Conf on Artificial Intelligence*, pp. 1044--1049.

Roberts A., Gaizauskas R., Hepple M. et al (2007). The CLEF Corpus: Semantic Annotation of Clinical Text. *AMIA 2007 Proceedings: Biomedical and Health Informatics: From Foundations to Applications to Policy*, pp. 625--629.

Roberts A., Gaizauskas R., Hepple M., Guo Y. (2008). Combining terminology models and statistical methods for entity recognition: an evaluation. *Proc Language Resources and Evaluation. Accepted for publication.*

Rogers J, Puleston C, Rector A. (2006). The CLEF chronicle: patient histories derived from electronic health records. *Proc 22nd Int Conf on Data Engineering Workshops.* p. 109.

Rosario, B. and Hearst, M. (2004). Classifying Semantic Relations in Bioscience Text. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL 2004).

Rosario, B. and Hearst, M. (2005). Multi-way Relation Classification: Application to Protein-Protein Interaction. *HLT-NAACL'05*, Vancouver.

Sager N, Lyman M, Bucknall C, Nhan N, Tick LJ. (1994). Natural language processing and the representation of clinical data. *J Am Med Inform Assoc.* 1(2), pp. 142--160.

Tanabe L., Xie N., Thom L., Matten W. and Wilbur J. (2005). GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics,* 6, suppl. 1:S3.

TimeML (2004). http://www.cs.brandeis.edu/~jamesp/-arda/ time/.

Verhagen, M., Gaizauskas. R., Schilder, F., Hepple M., Katz, G. and Pustejovsky, J. (2007). SemEval-2007 Task 15: TempEval Temporal Relation Identification. *Proceedings of the 4th International Workshop on Semantic Evaluations*, pp. 75-80.

Categorising Modality in Biomedical Texts

**Paul Thompson[1], Giulia Venturi[2,] John McNaught[1], Simonetta Montemagni[2]
and Sophia Ananiadou[1],**

[1]National Centre for Text Mining, University of Manchester, UK
[2]Istituto di Linguistica Computazionale, CNR (Italy)
E-mail: {paul.thompson, john.mcnaught,sophia.ananiadou}@manchester.ac.uk,
{giulia.venturi,simonetta.montemagni}@ilc.cnr.it

Abstract

The accurate recognition of modal information is vital for the correct interpretation of statements. In this paper, we report on the collection a list of words and phrases that express modal information in biomedical texts, and propose a categorisation scheme according to the type of information conveyed. We have performed a small pilot study through the annotation of 202 MEDLINE abstracts according to our proposed scheme. Our initial results suggest that modality in biomedical statements can be predicted fairly reliably though the presence of particular lexical items, together with a small amount of contextual information.

1. Introduction

Text processing systems tend to focus on factual language (Hahn & Wermter, 2006; McNaught & Black, 2006). However, modality is a common phenomenon which must be taken into account to correctly interpret text. Modality is concerned with the opinion and attitude of the speaker (Lyons, 1977). Palmer (1979) distinguishes three types of modality: *epistemic* (making judgements about the truth of a proposition), *deontic* (concerned with permission) and *dynamic* (concerned with the potential of a situation to occur).

Our concern here is *epistemic* modality in biomedical text, which covers the expression of the author's level of confidence towards a proposition, but may also indicate the type of knowledge, assumptions or evidence on which the proposition is based (Coates, 1995).

It also covers *speculation.* Light et al. (2004) and Medlock & Briscoe (2007) show that successful classification of biomedical sentences for speculation depends on the presence or absence of speculative cue words or phrases. Whilst more complex syntactic contexts, e.g. conditional clauses, are a possible way express modality in texts (Sauri et al, 2006), corpus-based studies of *hedging* (i.e. speculative statements) in biological texts by Hyland (1996a, 1996b) reinforce the above experimental findings: 85% of hedges were realised lexically, rather than through more complex means.

Previous efforts at annotating modal information in biomedical texts (e.g. Light et al., 2004; Wilbur et al. 2006; Medlock & Briscoe, 2007) have been at the sentence or sentence fragment level only, without explicit annotation of the modal cue words. Given the importance of these cue words, a list of modal lexical items used within the field, categorised according to the information they express, would be a useful resource for the automatic processing of biomedical texts.

Previous lists (e.g. Hyland, 1996a; Rizomiliti, 2006) suffer from being either incomplete or coarse-grained. Here, we describe the collection and multi-dimensional classification of a preliminary set of words and phrases that express modality within biomedical texts. We then report on initial validation of our classification, via annotation of modal information that modifies previously-annotated gene regulation events in a small corpus of MEDLINE abstracts.

Although oriented towards biological events, our proposed categorisation could be equally valid for other applications, e.g. helping to determine intent of citations, as suggested by DiMarco & Mercer (2004).

2. Modality in Scientific Texts

Hyland (1996a; 1996b) shows that modals such as *may, could, would* etc., play a relatively minor role in expressing modality in biological texts. This is evident when the proportions of word categories occurring in such texts are compared to those calculated by Holmes (1988) on general academic articles from the Brown corpus of American English and the LOB corpus of British English. See Table 1.

	Hyland (Biology)	Holmes (gen. academic)
Lexical verbs	27.4%	35.9 %
Adverbials	24.7%	12.8 %
Adjectives	22.1 %	6.6 %
Modal verbs	19.4 %	36.8 %
Nouns	6.4 %	7.7 %

Table 1: Comparison of modal items in different text types

It is the lexical (i.e. non-modal) verbs, adjectives and adverbs that dominate in expressing modality in biological research articles. Thus, we collected a set of modal words and phrases that are relevant within the biomedical domain.

3. Collecting Modal Items from Biomedical Texts

Rizomilioti (2006) provides a comprehensive base list of modal lexical items drawn from academic research articles in biology, archeology and literary criticism (200,000 words each).

As we hypothesised that modal lexical items can vary amongst text genres, we eliminated any items with no modal sense within the biomedical literature.

113 abstracts were taken from a corpus of MEDLINE abstracts on E. Coli (approximately 30,000 abstracts). Within these, any additional lexical markers of modality not present on Rizomilioti's list were identified. Examples included *evidence, observe, predict, imply, be consistent with* and *potential.*

For each item in the resulting combined list, we calculated its frequency within the complete E. Coli corpus, and discarded 26 items occurring fewer than 10 times (9 had zero occurrence). Discarded words included those indicating a high degree of confidence (e.g. *surely, patently, admittedly*, *attest, emphasise*) and those expressing doubt (e.g. *alledgedly, improbable, doubtful, ostensibly*).

We examined the usage of each remaining word in context within the corpus using a concordancer from the Multilingual Corpus Toolkit[1], and discarded any words not having a modal meaning in any contexts within the corpus. Examples included the verb *stress* and the adverb *naturally,* both of which Rizomilioti judged to indicate a high degree of confidence. In our corpus, *stress* almost always occurs as a noun, e.g. *oxidative/environmental stress*, whilst *naturally* most commonly occurs in phrases such as *naturally occurring.* Following this step, 90 lexical items remained. Table 2 shows the most frequently occurring of these in the E. Coli corpus, which correspond well with the highest ranked terms identified by Hyland and Rizomilioti.

show (17836)	may (5826)
suggest (11850)	demonstrate (4817)
indicate (8511)	reveal (4467)
observe (6177)	could (4247)
identify (5494)	appear (4212)

Table 2: Most frequent modal words in the E.Coli corpus

4. Classifying Modality in Scientific Texts

To propose a categorisation model for our list of modal lexical items, we considered a number of existing models and annotation schemes. Light et al. (2004) and Medlock & Briscoe (2007) are concerned primarily with the speculative/non-speculative distinction; other models are more complex and include multiple "dimensions". Rubin et al. (2005) annotate certainty in newspaper articles along 4 separate dimensions: a) degree of *certainty*; b) *focus,* i.e. whether the statement is abstract (opinions, beliefs, assessments) or factual; c) *perspective,* i.e. the writer's or a reported point of view; and d) *time,* i.e. whether the reported event is in the past, present or future. Lexical markers are explicitly annotated to give evidence for the value assigned to each attribute, suggesting that separate sets of words or phrases are used to express these different dimensions.

Although newspaper articles are very different from biomedical texts, the *certainty, focus* and *perspective* dimensions also seem relevant for us. The following corpus sentence illustrates the possibility of identifying these different dimensions through separate words or phrases:

We suggest that these two proteins may form a complex in the membrane which acts at late steps in the export process

The word *we* shows that the perspective is the authors' own, *suggest* provides *focus* information (i.e. this is a speculation rather than a definite fact) whilst *may* conveys the author's level of certainty.

4.1 Evidence Underlying Statements

An aspect not covered by Rubin et al.'s model, and yet highly relevant in scientific literature, is the source or type of *evidence* underlying a statement. The importance of this within the biomedical field is illustrated in annotation using the Gene Ontology (GO) (Ashburner et al, 2000). This requires gene associations to be attributed to the literature through the assignment of *evidence codes*, which denote the type of evidence available, e.g. experimental evidence, evidence through citations, or evidence inferred by the curator from other GO annotations.

Wilbur et al.'s annotation scheme also uses *evidence*; some of the possible values correspond closely to the main evidence categories used by GO curators, thus reinforcing that this type of information is important to domain experts. Sentences or sentence fragments are annotated for evidence as follows: a) no evidence, b) claim of evidence without verifying information, c) citation of other papers or d) explicit reference to experimental findings or results described within the paper.

4.2 Interpretation of Evidence

Certain verbs, like *see, indicate* and *find,* can help to identify statements containing reference to evidence. Wilbur et al.'s scheme determines the value of the *evidence* attribute largely from the type of subject taken by the verb, or the presence of citations. A subject such as *our results* provides explicit reference to results within the paper, whilst *previous studies* makes a claim of evidence, which may or may not be backed up by a citation.

Whilst the *type* of evidence behind a statement is important within the domain, another relevant type of information is how that evidence is to be *interpreted.* The choice of verb (or other modal lexical item) is important for this: whilst a statement beginning "*we **see** that …*" normally expresses an observation based on experimental findings or results, a sentence of the form "*Previous experiments **indicate** that…*" would imply that reasoning has taken place to arrive at the statement that follows.

Palmer's (1986) model for the 4-way classification of non-factual statements takes such distinctions into account, yielding *speculative, deductive* (derived from inferential reasoning or conclusions), *quotative* (specifying and acknowledging previous findings) and *sensory* (referring to apprehending, sensing or observing).

[1] http://personalpages.manchester.ac.uk/staff/scott.piao/research/

This model has been analysed for biological texts by Hyland (1996a). It has similarities with Wilbur et al.'s *evidence* attribute, in that the *quotative* category encompasses statements that cite other works. However, other types of statements backed by evidence are divided into *sensory* and *deductive,* according to whether they are based on observations or reasoning. Hyland's examples suggest that lexical items themselves can be used to distinguish between the *speculative, deductive* and *sensory* categories. For example, *appear* and *seem* are sensory verbs, whilst the verbs *propose, believe* and *speculate* fit well into the *speculative* category. Likewise, the verbs *infer, indicate* and *imply* are typical indicators of the *deductive* category.

5. Proposed Categorization Scheme

We conclude that the following factors are important to the interpretation of statements in scientific literature:

a) whether the statement is a speculation or based on factual data (e.g. experimental findings or results)
b) the type/source of the evidence
c) the interpretation of the evidence
d) the level of certainty towards the statement

We take Palmer's model as a starting point for our own proposed categorisation, as it covers the above factors a), b) and c), at least to a certain extent. Ad factor a), *sensory, deductive* and *quotative* statements are normally based on factual data, whilst speculations fall into the *speculative* category. Ad factor b), Palmer's categories allow different types of evidence to be distinguished. So, a *speculative* statement is not normally backed by evidence, whilst *sensory* and *deductive* statements would normally contain claims of evidence or reference to experimental findings. Meanwhile, *quotative* statements normally provide evidence through citation of other papers. Finally, ad factor c), different *interpretations* of evidence may be distinguished according to whether the statement is *sensory* or *deductive.*

Arguably, Palmer's categories implicitly encode certainty level information. A speculation is, for example, normally a less confident assertion than one backed by evidence of some sort. However, this does not necessarily follow: deductions and experimental observations may be made with varying degrees of confidence through the use of explicit certainty markers like *may* or *probably*. Thus, we follow Rubin et al. (2005) and Wilbur et al. (2005), in categorising certainty level as a separate dimension.

We further observed that *quotative* does not form a distinct category of statements. Consider: "*Trifonov [38] has suggested that…*". Here, the cited work speculates about the statement that follows, and so the sentence is *both* quotative *and* speculative. We thus classify the *point of view* of the statement (i.e. that of the author or a cited work) as a separate dimension. Whilst this does not correspond to modal information per se, its identification is important for correct interpretation of certain sentences containing modal lexical items, e.g. in determining the source of evidence presented. A sentence beginning O*ur data implies that* …is the author's point of view, indicating that the experimental findings discussed are drawn from the current paper rather than another source.

Our categorisation scheme for modality in biomedical texts thus consists of the following 3 "dimensions" of information:

1) *Knowledge Type,* encoding the type of "knowledge" that underlies a statement, encapsulating both whether the statement is a speculation or based on evidence *and* how the evidence is to be interpreted.
2) *Level of certainty,* indicating how certain the author (or cited author) is about the statement.
3) *Point of View*, indicating whether the statement is based on the author's own or a cited point of view or experimental findings.

Recognition of the *Point of View* level is aided through finding strings such as *we* and *our* (corresponding to the author's point of view), or various forms of citations for cited points of view. According to our scheme, the possible values for the *Point of View* dimension are *writer* or *other.* The other two dimensions can be recognised largely through the presence of lexical items such as the ones collected from our corpus of E. Coli abstracts.

5.1 Knowledge Type

The majority of lexical items within our list have been categorised under *Knowledge Type*. Three of the subclasses we use are taken from Palmer's model: *speculative, deductive* and *sensory.* To these, we add a fourth category of words whose members explicitly mark a statement as *describing* experimental results or findings, rather than observations or deductions made from them. Such statements are marked by words such as *show, reveal, demonstrate* or *confirmation.* As experiments are normally carried out to prove or demonstrate a hypothesis, we label this class of words *demonstrative.*

The largest category of items is the *speculative* one, containing 30 members from our preliminary list. These include not only verbs or their nominalised equivalents such as *predict, prediction, hypothesize, hypothesis,* etc., but also other nouns such as *view* and *notion,* adjectives like *conceivable* and phrases such as *in theory* and *to our knowledge.* Other categories are smaller, ranging from 8-12 items, consisting mainly of verbs and nominalised forms. So, a *deductive* statement can be denoted by *interpret, indication* or *deduce,* whilst sentences with *sensory* evidence can be marked with words such as *observation, see* or *appear.*

However, context may be required to correctly determine the category of statements denoted by certain *Knowledge Type* words: *suggest,* when used with a human subject, e.g. *We suggest …* or in the passive voice, e.g. *It is suggested…*, denotes a speculation; however, when the subject is inanimate, e.g. *The results suggest …,* there is an implication that a deduction has been carried out.

5.2 Level of Certainty

The partitioning of lexical items or statements into various degrees of certainty has been extensively studied, but little consensus has been reached. Rubin (2007) notes an ongoing discussion about whether they should be arranged on a continuum or into discrete categories.

Hoye (1997) proposes that there are at least three articulated points on the epistemic scale: *certainty,* *probability* and *possibility*. However, recent works have suggested more fine-grained partitions, with either 4 distinct levels (Rubin et al, 2005; Wilbur et al. 2006) or even 5 levels (Rubin, 2007). Annotation experiments according to this 5 level system (i.e. *absolute certainty, high certainty, moderate certainty, low certainty* and *uncertainty*) suggested that English may not be precise enough to distinguish so many shades between certainty and doubt. However, a 4-level distinction appears more feasible, with successful application in both the newspaper (Rubin et al., 2005) and biomedical domains (Wilbur et al., 2006). In the latter case, inter-annotator agreement rates of approximately 80% were reported.

Thus, we derived a four-way classification of lexical items denoting certainty: *Absolute*, *High*, *Moderate* and *Low.*

Within the scientific literature, a statement marked as *known* is normally an accepted fact within the field, and so is assigned the *Absolute* certainty value. Statements marked with words such as *probable, likely* or *clearly* express a *high* degree of confidence. Words such as *normally* and *generally* are also placed in this category, denoting that a specified event takes place most of the time, and thus expressing a high degree of confidence that the statement is true. Also within the *High* category are words and phrases that only express certainty when combined with certain *Knowledge Type* items. *Strongly* can be used in sentences of the following form: *The results strongly suggest that ….* Here, *suggest* has a *deductive* meaning, and *strongly* indicates a high degree of confidence towards this deduction. Words and phrases such as *support, in agreement with* and *consistent with* can be used with speculative nouns (e.g. *theory, notion* or *view*) to lower the speculation (and hence increase the certainty) of the statement.

Moderate items specify a more “neutral” certainty level, without strong indication of whether the statement is more likely to be true than false. Examples include *possibly* and *perhaps,* as well as some modal auxiliary verbs like *may* and *could.* Finally, *low* certainty level items have more negative undertone, signaling little confidence in the statement they modify, e.g. *unlikely*.

6. Testing the Classification Scheme

Our work has been carried out in the context of the BOOTStrep project (FP6 - 028099), aimed at building a bio-lexicon and bio-ontology for biomedical text mining. As part of the project, we have been creating a corpus of E. Coli abstracts annotated with *gene regulation* bio-events (Thompson et al., 2008). Events are centred around verbs (e.g. *regulate*) or nominalised verbs (e.g. *expression)*, and event annotation consists of identifying and classifying the semantic *arguments* or *participants* in the event. Note that event annotation was carried out on top of shallow parsed (pos-tagged and chunked) texts[2]: the advantages of such a choice range from practical ones, i.e. annotated corpora can be produced with much less work, to more substantial ones, i.e. previous levels of annotation can drive the annotation process, thus resulting in an increase in efficiency and consistency for any new annotation.

From the annotated events, patterns (i.e. *semantic frames*) relating to the behaviour of each verb and nominalised verb can be learnt and included within the bio-lexicon; these can help in the automatic extraction of facts from biomedical texts. As the annotated events correspond to facts of biomedical interest, we considered them a useful starting point for the verification of our proposed modality classification.

Thus, we carried out a small experiment, in which modality was annotated within a small set (i.e. 202) of these event-annotated abstracts, using *WordFreak,*a Java-based linguistic annotation tool (Morton & LaCivita, 2003), which was customized to the task.

Due to the linguistically-driven purposes as well as the small size of the corpus exploited in this feasibility study, annotation was carried out by a single annotator with linguistic expertise. However, extensive support was provided by two researchers, one with a background in linguistics, and the other one in biology, to discuss open issues raised during the annotation process in order to improve the semantic stability and reliability of the annotations produced.

6.1 Annotation process

Each sentence containing a previously-annotated gene regulation event was studied, and modality annotation was performed only on those sentences in which the description of the event contained explicit expression of modal information: modal information was only annotated if it was within the scope of the gene regulation event described. Let us consider, for example, the *derepress* bio-event, described in the sentence “*We suggest that overproduction of SlyA in hns(+) E. coli derepresses clyA transcription by counteracting H-NS”*, which was annotated as follows:

VERB: *derepresses*
AGENT: *overproduction*
THEME: *clyA transcription*
MANNER: *counteracting*

The modality annotation process started from the event anchor, i.e. the verb *derepress*. Words or phrases expressing modal information and linguistically bound to the event anchor were searched for within the sentence’s span. If such items were found, values from the proposed sets were selected for one or more of the three dimensions of the annotation scheme, i.e. *Point of View, Knowledge Type* and *Certainty Level*. For the *Knowledge Type* and *Certainty Level* attributes, a value was only selected if there was *explicit* lexical evidence in the sentence. In the case at hand, *suggest* was annotated as the lexical modality marker conveying information about *Knowledge Type*, whose associated value is *deductive*. The word *We* was interpreted as lexical evidence that the reported *Point Of View* was that of the writer.

Each piece of lexical evidence (i.e. lexical modality marker) could only be used to assign a value to *one* of the annotation dimensions. Thus, it was not possible to use a single word or phrase to assign values to both the *Knowledge Type* and *Certainty Level* dimensions.

If one or both the *Knowledge Type* or *Certainty Level* attributes were assigned, the *Point of View* attribute was

[2] Each abstract to be annotated with gene regulation bio-events was first pre-processed with the GENIA tagger (Tsuruoka et al, 2005).

also instantiated. If no explicit lexical evidence was available for the assignment of this attribute, a “default” value of *writer* was assigned, i.e. it was assumed that the *Point of View* was expressed implicitly.

The annotator used the preliminary categorisation of modal lexical items as a starting point for the annotation of the *Knowledge Type* and *Certainty Level* attributes, although she was not bound by this categorisation, nor was her annotation limited to only those items on the list: part of the purpose of the annotation was to discover the semantic stability of the lexical items within our proposed categories, as well as to discover other modality markers missing from the preliminary list.

7. Results

The 202 MEDLINE abstracts annotated for modal information contained a total of 1469 gene regulation events. 249 of these events (i.e. 16.95%) were annotated with modality information. Table 3 shows general statistics about the *dimensions* of the modal markers that were present in the description these events, whilst Table 4 shows the distribution of the annotations amongst the various values within each dimension of the scheme.

The number of modality annotations may at first seem rather low, with an average of 1.31 annotations per abstract. However, a number of points should be noted. Firstly, lexical markers of modality are generally quite sparse within texts. Secondly, as pointed out above, modality annotations have only been carried out on top of previously annotated bio-events, and there was an average of 6.05 bio-event annotations per abstract. Rather than aiming to annotate *all* modal information expressed within the abstracts, our case study is firstly aimed at verifying whether the modality classification scheme is suitable for a corpus of biomedical texts, and secondly, it is focused on the discovery of the main domain-relevant problems and features involved, as well as clues which can drive future work.

There follows a number of annotation examples. In each case, the modality marker(s) and the Point Of View marker (if present) have been underlined, with the corresponding category placed in brackets. The verb which forms the focus of the associated bio-event is emboldened.

a) *Therefore, we* [WRITER] *suggest* [DEDUCTIVE] *that overproduction of SlyA in hns(+) E. coli* ***derepresses*** *clyA transcription by counteracting H-NS.*

b)*We* [WRITER] *have shown* [DEMONSTRATIVE] *that the open reading frame ybbI in the genomic sequence of Escherichia coli K-12* ***encodes*** *the regulator of expression of the copper-exporting ATPase, CopA*

c)*We* [WRITER] *speculate* [SPECULATIVE] *that the product of this gene is* ***involved*** *in the attachment of phosphate or phosphorylethanolamine to the core and that it is the lack of one of these substituents which results in the deep rough phenotype.*

A single modality marker may also express the same information relative to more than one bio-event in the case of a coordinated structure, e.g. :

Band shift experiments showed [DEMONSTRATIVE] *that AllR* ***binds*** *to DNA containing the allS-allA intergenic region and the gcl(P) promoter and its binding is* ***abolished*** *by glyoxylate.*

Modal marker(s) present	**Count**	**% of total events**
Knowledge Type only	192	77.11 %
Certainty Level only	40	16.07%
Knowledge Type + Certainty Level	17	6.83%

Table 3: Distribution of modality markers within annotated events

Dimension	**Value**	**Count**	**% of annotations within dimension**
Knowledge Type	DEMONSTRATIVE	110	52.63%
	DEDUCTIVE	56	26.79%
	SENSORY	25	11.96%
	SPECULATIVE	18	8.61%
Certainty Level	ABSOLUTE	4	7.01%
	HIGH	15	26.31%
	MODERATE	34	59.64%
	LOW	2	3.50%
Point Of View	WRITER	213	92.20%
	OTHER	18	7.79%

Table 4: Distribution of modality annotations within the different dimensions

7.1 *Knowledge Type* Information

The *Knowledge Type* dimension is the most frequently annotated (77.11% of annotations). The most common value for this dimension is *demonstrative* (52.63% of Knowledge Type annotations), whilst the least widespread type of knowledge is *speculative* (8.61%).

These statistics are perhaps unsurprising, given that the current pilot study has been carried out on abstracts. *Demonstrative* events are explicitly marked as describing experimental results, particularly those which prove hypotheses or predictions. These are exactly the sorts of events that we can expect to occur most frequently in abstracts; within the short amount of space available, authors normally aim to emphasize the definite results that their experiments have produced.

The annotation experiment has highlighted a potential need to add an additional value for the *Knowledge Type* dimension. Consider the following examples:

a) *The model states that the lex (or exrA in E. coli B) gene* ***codes*** *for a repressor.*

b) *Mutations in yjfQ allowed us to identify this gene as the regulator of the operon yjfS-X (ula operon), reported to be* ***involved*** *in L-ascorbate metabolism.*

Events that are introduced by verbs such as *state* or *report* do not fit well into one of our other four *Knowledge Type* categories. They are used to introduce facts, either cited from previous work or earlier in the paper, but without taking a particular stance to them, i.e. there is no speculation or deduction involved, and there is no reference to active proof or demonstration that an assertion or hypothesis is true.

Statements such as the above fit into Hyland's (1996a) description of the *quotative* category, i.e. specifying and acknowledging previous findings. Thus, the *quotative* label can apply to a wider range of statements than just those that contain citations. Therefore, we propose to introduce the *quotative* category into our classification as a further *Knowledge Type* category to cover statements that specify or acknowledge previous findings through explicit lexical items.

Our annotation also revealed that, whilst the majority of *Knowledge Type* items are fairly stable semantically within their assigned categories, a small number of items do not fit neatly within a single category. The verb *seem* was originally placed within the *sensory* category, following Hyland. However, there is often a speculative aspect to its meaning, as confirmed by Dixon (2005): *seem* is used "when there is not quite enough evidence" (p. 205). The degree of speculation conveyed may vary according to the context: this is an area for further research.

7.2 *Certainty* Information

Certainty level markers are considerably less common than *Knowledge Type* markers, representing 16.07% of the modality annotations. The most widespread value among these annotations is *moderate* (59.64%).

The high percentage of *moderate* markers can again be explained by the text type, i.e. abstracts. The results concerning *Knowledge Type* illustrated that *demonstrative* statements are most common: authors are keen to emphasize the experimental results that they have produced. If there is doubt about these results, this can be indicated thought an explicit certainty level marker. A *moderate* (and hence neutral) certainty level marker may be the "safest" choice here.

Certainty Level markers occur most commonly without an accompanying *Knoweldge Type* marker, as in:

EvgA is likely [HIGH] to directly ***upregulate*** *operons in the first class, and indirectly upregulate operons in the second class via YdeO.*

As mentioned previously, *Knowledge Type* markers implicitly encode certainty level information. Thus, when a statement is explicitly marked as a speculation or deduction, the use of an explicit marker of certainty may be unnecessary, except for emphasis, or to alter the "default" certainty level associated with the *Knowledge Type* item.

Nevertheless, our annotation has served to identify a small number of cases (6.83%) that contain explicit markers of both *Knowledge Type* and *Certainty Level* information. Such cases provide evidence that our proposed separate dimensions of annotation are indeed well motivated. Some examples are shown below:

a) *No reverse transcriptase PCR product could be detected for hyfJ-hyfR, suggesting* [DEDUCTIVE] *that hyfR-focB may* [MODERATE] *be independently* ***transcribed*** *from the rest of the hyf operon.*

b) *We* [WRITER] *suggest* [SPECULATIVE] *that these two proteins may* [MODERATE] ***form*** *a complex in the membrane which acts at late steps in the export process.*

A large number of certainty level markers are fairly stable in terms of semantics, particularly adjectives and adverbs such as *probable, possibly* or *likely.* Another category of words that play a central role in expressing certainty in our corpus is the modal auxiliaries (e.g. *can, may* or *could*), which represent 40.35% of the total number of *Certainty Level* markers. However, their interpretation is more problematic than adjectives and adverbs like those listed above. In general, *can, may* and *could* can have the following senses:

1) *Moderate* level of certainty
2) Theoretical possibility (indicating that an event has the potential to occur)
3) Ability
4) Permission

Whilst the *permission* sense is rarely relevant within biomedical texts, examples of the other three senses can be readily identified within our corpus. Some examples involving *may* are shown below:

1) **Certainty level marker**
The DNA-binding properties of mutations at positions 849 and 668 may [MODERATE] *indicate* [DEDUCTIVE] *that the catalytic role of these side chains is* ***associated*** *with their interaction with the DNA substrate.*
2) **Theoretical possibility marker**
The expression of nifC may be ***coregulated*** *with nitrogen fixation because of the presence of nif-distinctive promoter and upstream sequences preceding nifC-nifV omega-nifV alpha.*
3) **Ability marker**
Results obtained indicate that the nrdB gene has a promoter from which it may be ***transcribed*** *independently of the nrdA gene.*

Thus, the presence of these modal auxiliaries does not guarantee that certainty level is being conveyed. Determining the correct sense can be a difficult task, which requires in-depth knowledge of the domain, and often requires examining a wider context than just the sentence itself.

Whilst this could prove problematic in the automatic recognition of modality, Collins (2006) suggests that for each verb, one sense is usually more likely than the others. In his study of *can* and *may* in various spoken and written sources, he found that *may* was used as a certainty level marker in 83.5% of cases, whilst only 1.1% of occurrences of *can* concerned certainty level. A default interpretation of each modal could thus be used. Further study of the context of these items may reveal clues that could determine when a non-default value should be assigned.

Our studies have shown that the meaning of *can* mainly corresponds to the "ability" sense, although "theoretical possibility" is also possible, as shown in the following examples:

a) *The enhanced expression of tac-dnaQ reduces 10-fold the frequency of UV-induced Su+ (GAG) mutations in the CCC phage and nearly completely prevents generation by UV of Su+ (GAG) mutations in the GGG*

phage, in which UV-induced pyrimidine photo-products can *be* ***formed*** *only in the vicinity of the target triplet.*
b) *These results indicate that OmpR stabilizes the formation of an RNA polymerase-promoter complex, possibly a closed promoter complex, and that a transcription activator* can ***serve*** *not only as a positive but also as a negative regulator for gene expression.*

Whilst the "ability" sense is not central to the interpretation of modality, the recognition of "theoretical possibility" may be more important: stating that an event has the *potential* to happen is different from stating that it *does* (always) happen. Thus, further investigation of lexical markers of theoretical possibility will help to build upon our current categorisation model.

7.3 *Point Of View* Information

Although we suggested that there are a number of textual clues that can be used to determine the *Point of View* of a statement, our annotation experiment revealed that such explicit evidence is quite sparse, at least in abstracts. Occasionally, the sentence contains words or phrases such as *we*, *our results*, *in this study*, etc. allowing the *Point Of View* to be determined as the author(s) of the abstract. In other cases, looking at the wider surrounding context, i.e. in neighbouring sentences or even within the whole abstract, is necessary. Although our annotation assumes the lack of an explicit *Point of View* marker to indicate the *writer* point of view, further analysis of these cases must be carried out.

During annotation, however, we identified some potential additional clues that can help to determine the value of this dimension.

Consider the phrase *these results*. On its own, this provides no explicit information about the point of view of the accompanying statement. However, when occurring as the subject of *suggest* (especially in the present tense), it is normally the case that the deduction has been carried out by the author(s), as illustrated in the following example:

These results [WRITER] suggest [DEDUCTIVE] *that both locally and regionally targeted mutagenesis is* ***affected*** *by overproduction of the epsilon subunit.*

The *writer* value can also be assumed in such contexts when other verbs in the *deductive* and *sensory* categories are used, e.g. *indicate, imply, appear*, etc, particularly when in the present tense with an inanimate subject. An exception is when there is explicit reference to another author or work. If there is an impersonal subject, e.g. *It is suggested*, then greater contextual evidence would be required, as the point of view is ambiguous.

A further example concerns *Certainty Level* markers within the *absolute* category, which generally denote well-established facts within the community. When such a certainty level marker is present, we can assume that the statement does not correspond only to the author's personal point of view. An example is shown below:

Near the amino terminus is the sequence 35GLSGSGKS, which exemplifies a motif known [ABSOLUTE] *to* ***interact*** *with the beta-phosphoryl group of purine nucleotides.*

8. Conclusion

We have presented a scheme for classifying modality in biomedical texts according to three different dimensions, namely *Knowledge Type, Certainty Level* and *Point of View.* In many cases, textual clues can be used fairly reliably to determine the correct classification of statements according to these dimensions. The results from a preliminary annotation experiment based on this scheme confirm this hypothesis.

Contextual information surrounding modal lexical items can also be important in determining the correct modal value of statements. Shallow parsing (i.e. chunking), on the top of which event annotation and modality annotation are carried out, can help to identify such information. This is in agreement with Medlock & Briscoe (2006), who suggest that linguistically-motivated knowledge may help to boost the performance of an automatic hedge classification system.

Our preliminary results suggest that many modal items in our list are fairy stable semantically when modifying bio-events. However, the correct interpretation of modal auxiliaries within the domain is more problematic, and is thus an area for further research. Our experiment also served to highlight certain weaknesses in the original model, e.g. the lack of a *Knowledge Type* category corresponding to reported facts. A further potential weakness in our results is that, whilst examples supporting all of our proposed categories were found, there is a strong bias towards certain categories. This may be because our preliminary study was based only on abstracts.

In the future, we plan to carry out further experiments to reinforce the validity of our proposed classification. These include involving multiple annotators (including biologists) to provide inter-annotator agreement statistics, as well as applying our scheme to full texts, where we can expect a greater variability of modal expression to be encountered.

9. Acknowledgements

The work described in this paper has been funded by the European BOOTStrep project (FP6 - 028099). We would also like to thank Philip Cotter for his generous help with biomedical issues.

10. References

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. & Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetetics,* 25, pp 25--29.

Coates, J. (1995). The expression of root and epistemic possibility in English. In B. Aarts & C. F. Meyer (Eds.), *The Verb in Contemporary English. Theory and Description.* Cambridge: Cambridge University Press, pp 145--156.

Collins, P. C. (2006). Can and may: monosemy or polysemy?. In I. Mushin & M. Laughren, (Eds.), *Annual Meeting of the Australian Linguistic Society*, Brisbane, Australia.

DiMarco, C., & Mercer, R.E. (2004). Hedging in scientific articles as a means of classifying citations. In *Working Notes of the American Association for Artificial Intelligence (AAAI) Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, pp 50--54.

Dixon, R. M. W. (2005) *A Semantic Approach to English Grammar*. Oxford: Oxford University Press.

Hahn, U. & Wermter, J. (2006). Levels of Natural Language Processing for Text Mining. In S. Anananiadou & J. McNaught (Eds.), *Text Mining for Biology and Biomedicine.* London: Artech House, pp. 13--42.

Holmes, J. (1988). Doubt and certainty in ESL textbooks. *Applied Linguistics,* 13(2), Oxford: Oxford University Press, pp. 21--44.

Hoye, L. (1997). *Adverbs and Modality in English.* London, New York: Longman.

Hyland, K. (1996a). Talking to the Academy: Forms of Hedging in Science Research Articles. *Written Communication*, 13(2), pp.251--281.

Hyland, K. (1996b). Writing Without Conviction? Hedging in Science Research Articles. *Applied Linguistics* 17(4), Oxford: Oxford University Press, pp. 433--454.

Light, M., Qiu, X.Y. & Srinivasan, P. (2004). The Language of Bioscience: Facts, Speculations, and Statements In Between. In *Proceedings of BioLink 2004 Workshop on Linking Biological Literature, Ontologies and Databases: Tools for Users*, pp.17--24.

McNaught, J & Black, W. (2006). Information Extraction. In S. Ananiadou & J. McNaught (Eds), *Text Mining for Biology and Biomedicine.* London: Artech House, pp. 143--178.

Medlock, B. & Briscoe, T. (2007). Weakly Supervised Learning for Hedge Classification in Scientific Literature. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, Czech Republic, pp. 992--999.

Morton T. & LaCivita J. (2003). Word-Freak: an open tool for linguistic annotation. In *Proceedings of HLT/NAACL-2003*, pp. 17--18.

Palmer, F. (1986). *Mood and modality.* Cambridge: Cambridge University Press

Rizomilioti, V. (2006). Exploring Epistemic Modality in Academic Discourse Using Corpora. *Information Technology in Languages for Specific Purposes* (7), pp. 53--71.

Rubin, V. (2007). Stating with Certainty or Stating with Doubt: Intercoder Reliability Results for Manual Annotation of Epistemically Modalized Statements. In *Proceedings of The Human Language Technologies Conference: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*, *Companion Volume*, pp. 141--144.

Rubin, V. L., Liddy, E. D., & Kando, N. (2005). Certainty Identification in Texts: Categorization Model and Manual Tagging Results. In J. G. Shanahan, Y. Qu & J. Wiebe (Eds.), *Computing Attitude and Affect in Text: Theory and Applications (the Information Retrieval Series*. New York: Springer-Verlag, pp. 61--76.

Sauri, R., Verhagen, M., & Pustejovsky, J. (2006). Annotating and Recognizing Event Modality in Text. In *Proceedings of the 19th International FLAIRS Conference, FLAIRS 2006*. Melbourne Beach, Florida, pp. 333--339.

Thompson, P., Cotter, P., Ananiadou, S., McNaught, J., Montemagni, S., Trabucco, A., Venturi, G., (2008). Building a Bio-Event Annotated Corpus fro the Acquisition of Semantic Frames from Biomedical Corpora, to appear in *Proceedings of Sixth International Conference on Language Resource and Evaluation (LREC 2008).*

Tsuruoka, Y., Tateishi, Y., Kim, J-D., Ohta, T., McNaught, J., Ananiadou, S. & Tsujii, J. (2005). Developing a Robust Part-of-Speech Tagger for Biomedical Text, In *Advances in Informatics - 10th Panhellenic Conference on Informatics*, pp 382--392.

Wilbur, W.J., Rzhetsky, A. and Shatkay, H. (2006) New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinformatics* 7:356

Appendix A: Lexical Modality Markers

Knowledge Type Markers

Speculative - *assume, assumption, belief, believe, claim, conceivable, estimate, expect, expectation, hypothesise, hypothesis, hypothetical, in principle, in theory, judge, model, notion, predict, prediction, proposal, propose, speculate, suggest[3], suggestion, suppose, suspect, theory, think, to our knowledge, view.*

Deductive – *argue, argument, deduce, imply, indicate, indication, infer, interpret, interpretation, suggest[4].*

Demonstrative - *conclude, conclusion, confirm, confirmation, demonstrate, find, finding, proof, prove, report, reveal, show.*

Sensory - *apparent, apparently, appear, observation, observe, evidence, evident, seem, see.*

Certainty markers

Absolute - *certainly, known.*

High - *consistent with[5], clear, clearly, generally, in agreement with[5], likelihood, likely, normally, obviously, probability, probable, probably, strongly[6], support[5], would.*

Medium – *can, could, feasible, may, might, perhaps, possibility, possible, potential, potentially.*

Low – *unlikely, unknown.*

[3] with a human subject, e.g. *We suggest that* … or in the passive voice, e.g. *It is suggested that…*

[4] with an inanimate subject, e.g. *The results suggest that …*

[5] Often used to lower the speculation (and hence increase the certainty) of a speculative statement, e.g. *These results are consistent with the view that …*

[6] Often used to strengthen the certainty of deductive or speculative propositions, *e.g. The results strongly suggest that …*

Static Dictionary Features for Term Polysemy Identification

Piotr Pęzik, Antonio Jimeno, Vivan Lee, Dietrich Rebholz-Schuhmann
European Bioinformatics Institute
Wellcome Trust Genome Campus
Hinxton, Cambridge, CB10 1SD, UK
E-mail: piotr.pezik@gmail.com

Abstract

Building a large lexical resource which pools together terms of different semantic types naturally leads to the consideration of term ambiguity issues. Both cross- and intra-domain term polysemy constitutes a formidable obstacle in tasks and applications such as Named Entity Recognition or Information Retrieval (e.g. query expansion or relevance feedback), where a single polysemous term may cause a significant concept drift. One of the biggest sources of polysemy in biomedical terminologies are protein and gene names (PGN), both those extracted from existing databases and their variants found in the literature. We provide an analysis where the effect of using static dictionary features for the detection of potential polysemy of protein and gene names (and by extension other semantic types) is clearly delineated from the contribution of applying contextual features. We argue that, although disambiguation based on static dictionary features does not outperform fully-fledged context-driven Named Entity Recognition, it does effectively filter out highly polysemous terms (increase in F-measure from 0.06 to 0.57 and from 0.21 to 0.51 as measured on two evaluation corpora). Moreover, static dictionary features are context-independent and thus more easily applicable in systems where running intense on-the-fly disambiguation for retrieved documents could be problematic.

1. Introduction

The preparation and refinement of a lexical resource constitutes an important step in many information extraction and information retrieval systems used for the biomedical domain. Unfortunately, very few of the widely available databases seem to have all the makings of a high-quality lexical resource. As an example, ontologies and taxonomies (e.g. GO, or MESH) come with little or no guarantee that the potential terms they contain are actually used in the relevant literature. Resources such as BioThesaurus [1] contain a mixture of terms and pseudo-terms extracted semi-automatically from a range of varying quality protein and gene databases. This leads to a high level of both artificial and genuine term ambiguity, which can be dealt with both at the level of the dictionary, and the context of term occurrence. The need for disambiguation techniques is recognized in Named Entity Recognition approaches, which, within the biomedical domain, have been mainly focused on the identification of protein and gene names (PGNs)[2]. This focus is largely due to the inherent polysemy of PGNs as well as their importance as one of the fundamental entities of interest in the biomedical domain, although other semantic types such as enzymes, chemical entities and species also have to be handled in text mining applications. The recognition and normalisation of protein and gene names has received much focus in community-wide Information Extraction evaluation efforts such as BioCreAtivE[3]. The issue of term polysemy is also important in the field of Information Retrieval (i. e. query expansion or relevance feedback), for the biomedical domain, where a single polysemous term may cause a significant concept drift (e.g. Jimeno et al. 2007). In this paper we first introduce a terminological resource combining a number of semantic types imported from different biomedical databases. We show some potential advantages of having different semantic types pooled together, and how the intrinsic polysemy of such a combined resource can be dealt with. By a semantic type we mean a distinct class of biological or chemical entities, e.g. protein names as opposed to species names. Next, we focus on the task of automatic protein/gene names recognition to demonstrate how the performance of a named entity recognition system relying on static information encoded in a lexical resource can be significantly improved. Having analyzed the different usage patterns of E. coli and human PGNs, we have identified four types of polysemy affecting PGNs stored in our lexical resource and implemented a number of dictionary-filtering rules addressing each of these types as a set of decision tree features. As a result, we manage to improve the performance of a noisy dictionary-based protein name recognition system. Our experiments confirm that static dictionary features contribute significantly to the identification of highly polysemous terms. At the same time we attempt to systematize the major types of PGN polysemy and clearly delineate the influence of static dictionary features from text driven ones.

2. Term Repository

Table 1 shows the contents of the term repository we have used to test static dictionary features.

[1] http://pir.georgetown.edu/pirwww/iprolink/biothesaurus.shtml. We used BioThesaurus v. 2 for the experiment reported here. BioThesaurus only contains protein and gene names mapped to Uniprot accession numbers. We have taken advantage of this mapping and used the Uniprot annotations such as the species identifiers available for each PGN.

[2] Many NER techniques deal with resolving the ambiguity of polysemous terms and in achieving this goal they derive from the field of Word Sense Disambiguation.

[3] http://biocreative.sourceforge.net/

Semantic Type	Synsets	Variants
CHEBI	13,473	57,581
Enzyme names	4,016	7,658
PGNs	232,258	1,931,786
Species names	367,565	441,993

Table 1 Synsets and variants in the Term Repository

Currently, the Term Repository integrates terms from different resources, including a subset of Biothesaurus (Liu et al. 2006) protein/gene names (PGNs), which as demonstrated below ensures a relatively high initial recall of PGNs. The repository also contains Chemical Terms of Biological Interest (CHEBI[4]), enzyme names[5], and species names imported from the NCBI species taxonomy[6]. All of these terms are organized into clusters or synsets, i.e. sets of term variants which can be mapped to the same original database sense identifier.

Although the biggest number of distinct senses is contributed by the NCBI species taxonomy, the biggest number of synonyms is introduced by PGNs. Depending on the semantic type, it is possible to annotate each cluster and each term with metadata derived from the Term Repository itself or from external resources. As an example, each orthographically distinct PGN in the repository can be annotated with relevant species identifiers, frequencies in reference corpora or the number of nodes where it appears in external references. We show how such metadata can be used to tackle the problem of term polysemy.

The PGNs imported from the Biothesaurus come from a variety of different resources. In theory, a PGN recognition tool using a comprehensive terminological resource holds a promise of a high recall value (Schuemie et al. 2007), at the expense of precision due to a higher level of polysemy introduced by orthologous gene names, as well as noisy terms coinciding with PGNs.

As a first step in the preparation of the Term Repository for a PGN recognition task, we removed the most obvious pseudo terms such as *hypothetical protein*, *putative protein* or *possible protein* and terms corrupted in the process of automatic integration. Needless to say, such a filtering method does not deal with truly polysemous terms. After this initial clean-up and some basic case normalization, we used the Term Repository as a dictionary of PGNs.

PGNs tagged as human relevant in the repository based on their Uniprot species annotation were selected from and compiled into a deterministic final state automaton-based regular expression engine. Although the recall was relatively high (0.82), the precision of 0.03 obtained for the gene normalization task on the BioCreAtivE corpus was unacceptably low. To increase precision, we first carried out a detailed analysis of the types of polysemy found in PGN name resources. For the four major types of polysemy identified, we designed six features providing indications of one or more types of PGN polysemy. While defining the features we made sure that they are either derived directly from the Term Repository itself, or assigned statically to each term, and thus easy to use for other semantic types and other terminological resources, since they do not rely on the morphology of terms. As reported in the results section of this paper, these text-independent features significantly increased the overall performance of PGN recognition.

A purely dictionary-based method cannot compete with the performance of state-of-the-art approaches which combine dictionary look-up with contextual features. Therefore, for the sake of comparison, we complement the dictionary filtering rules with a set of contextual features. It has to be stressed however, that the main focus of the experiment reported here lies in the definition of resource-independent dictionary filtering rules which address the four types of PGN polysemy identified. This is meant to address the recently recognized need for a systematic feature evaluation in PGN identification systems (Hakenberg et al. 2005). The analysis we present below depends on a clear mapping between the rules and features applied and the types of polysemy they address. Both sets of features were fed into a decision tree-based model and evaluated against the BioCreAtivE corpus. We report on the model's performance for dictionary and text features used independently of each other.

3. Major types of PGN polysemy

Having analyzed the PGNs imported from the Biothesaurus into the Term Repository, we distinguish between the following cases of genuine PGN polysemy:

1. A PGN has a common English word homograph. We call this a case of domain-independent polysemy, e.g. (but, WHO). Sometimes this type of polysemy is introduced by pseudo terms by resulting from the poor quality of a lexical resource, e.g. Biothesaurus contains partial PGN terms such as human or, due to the fact that they were gathered from less trustworthy database description fields.
2. A PGN has a number of hyponyms and it is sometimes used synonymously with them. Examples of this type of polysemy include generic enzyme names, such as *oxidoreductase*). Sometimes a more specified case of holonymy triggers similar ambiguity, e.g. an operon name can be interpreted to denote any of the genes it contains. We call this a case of vertical polysemy (c.f. Fellbaum 1998).
3. A PGN is used for a number of orthologous or otherwise homologous genes. Thus the ambiguity in the gene name results from the fact that the same

[4] http://www.ebi.ac.uk/chebi/

[5] http://www.chem.qmul.ac.uk/iubmb/enzyme/

[6] http://www.ncbi.nlm.nih.gov/sites/entrez?db=taxonomy

name is used for structurally identical genes found in different species.

4. A PGN has a biomedical homograph, e.g. retinoblastoma. We refer to this as a case of domain-specific polysemy (Jimeno et al. 2008).

Last but not least the very use of the umbrella term PGN suggests another type of polysemy, where the same name is used to denote a gene and its product. Generally, however, gene names are not distinguished from protein names.

Type 2 and Type 3 polysemy have also been described as intra- and cross-species ambiguity (Chen et al. 2004).

4. Feature set for polysemy detection

4.1 Static dictionary features

It is possible to design a feature set for these four types of polysemy introduced above. Table 2 shows 6 features and the corresponding types of polysemy they may indicate.

#	Feature	Polysemy type	Data type
1	BNC frequency	1	Integer
2	Number of synsets	2,3	Integer
3	NCBI taxonomy ids	3	Integer
4	Generic enzyme	2	Boolean
5	Medline frequency	4,1	Integer
6	MESH nodes	4	Integer

Table 2 Static dictionary feature set addressing the four types of polysemy.

BNC frequency is the frequency of a given PGN in the British National Corpus and it is meant to give a clue about possible do-main-independent polysemy (Type 1 polysemy).

As explained above, similarly to Biothesaurus, the Term Repository organizes terms into sets of synonymous PGNs (clusters), which can be mapped to the same UniProt accession number. *Feature 2* is thus meant to provide indications of Type 2 and Type 3 polysemy at the same time; the fact that a PGN is found under different accession numbers may mean that it is used for different orthologous genes and/or that it is a generic name which can be used to refer to many different more specific PGNs.

The *number of NCBI species taxonomy identifiers* assigned to the PGN provides evidence of Type 3 polysemy.

For feature 4, we have manually tagged 37 PGNs as *highly polysemous generic enzyme names* (e.g. oxidoreductase). This Boolean flag gives indications of Type 2 polysemy.

Feature 5 is the *frequency of a PGN in Medline*. This value is meant to provide information on Type 4 polysemy, although it does contain some information on Type 1 polysemy as well. A combination of World Street Journal and Medline frequencies has been used for PGN disambiguation by (Tanabe & Wilbur 2002).

Feature 6 is the number of times a term occurs under *non-protein MESH nodes*. This feature encodes possible indications of Type 4 polysemy.

In addition to these features, we decided to break down PGNs into their constituent word tokens. This feature does some justice to the morphological characteristics of potentially polysemous PGNs. It is possible to relate it to all of the four types of polysemy specified above, but this relation is rather latent and PGN name dependent.

We assigned each of these features to a set of terms extracted from the training data sets, labeling each as a true positive or a true negative. The annotation of PGNs in the BioCreAtivE corpus is provided at the level of a span, while in the E. coli corpus the exact strings denoting PGNs are marked-up. Any terms annotated by the dictionary look-up method as PGNs, but not in the training section of the BioCreAtivE corpus were considered as true negatives. All the other identified PGNs were considered to be true positives.

String matching techniques are often used to enable fuzzy dictionary look-up in order to take advantage of the regularities in the morphology of PGNs which may be indicative of its polysemy. (Wilbur et al. 2007). For the purposes of this experiment we do only very basic case and hyphen normalisation. Some more extensive work on fuzzy look-ups which may potentially increase the recall is in progress.

We fed the training sets into the Weka machine learning package implementation of the C4.5 decision tree algorithm (Quinlan, 1993). The different parameters for the algorithm were estimated for each dataset via 10-fold cross-validation. We show the strongest rules identified by the decision tree in the following sections.

4.2 Text-driven features

We have so far identified a number of generic rules supporting PGN recognition that can be derived from a dictionary and other static resources, and which can be relatively easily applied to other semantic types and resources. The results yielded by such dictionary filtering rules alone produce a significant improvement in the F-measure. However, they are still below the results reported for complete PGN normalization approaches evaluated on the BioCreAtivE corpus, which use both lexicon filtering and text-driven rules, with reported results in the region of 0.7-0.8 F-measure (Morgan, Hirschman 2007). To demonstrate that we have also tested the improvement gained when the dictionary filtering feature vector is complemented with a text-driven one. For that purpose we have adopted the following six corpus-specific contextual features for PGN identification.

1. The frequency of the term in the BNC Corpus.
2. The frequency of the term in Medline.
3. Whether the term is annotated with more than one identifier.
4. The frequency of the term in a given passage.
5. The number of distinct terms linked to the same term in the passage.
6. Whether the term matches the boundaries of an entity identified (but not normalized) by Abner (Settles, 2005), a conditional random field-based named entity extraction tool that has shown good performance on BioCreAtivE and NLPA datasets (F-measure of 0.69 and 0.705 respectively).

A decision tree trained on this set of contextual features is applied on top of the dictionary-based approach. The improvements obtained are reported in the following sections.

5. Evaluation corpora

We have evaluated the use of static dictionary features in an NER scenario, as it is directly relevant to the problem of ambiguity resolution. The primary evaluation of the dictionary-based disambiguation features is based on the corpus used for the BioCreAtivE gene normalization task, in which abstracts have to be annotated with EntrezGene identifiers. The BioCreAtivE dataset contains 281 manually annotated abstracts together with some 5000 documents from the Gene Ontology Annotation database. The latter part of the corpus is used as noisy training data. There also are 252 abstracts provided in the test set. However, because the Term Repository contains PGNs covering other species, we thought it necessary to carry out the evaluation of the dictionary-based PGN disambiguation features on a corpus representative of non-human PGNs as well. For that purpose we have used a PGN-annotated corpus currently containing 96 abstracts related to the topic of gene regulation in E. coli. E. coli is one of the prokaryotic model organisms which is not covered by any of the BioCreAtivE data sets, so we thought it useful to perform the method described here on a novel corpus covering this species. The minimum requirement for pre-selecting an abstract was the occurrence of at least one string matching an E. coli PGN recorded in SwissProt and some keywords indicating E. coli as a species. This resulted in a set of 33,635 abstracts of which 96 abstracts relevant to the topic of gene regulation were extracted and manually verified. One important difference between this corpus and the BioCreAtivE one is that it annotates the exact occurrences of a PGN (rather than a span of text where it occurs) and assigns possibly more than one UniProt accession numbers to the PGN if it is truly ambiguous in the text (e.g. porin). A successful recognition of a protein is only counted if the exact boundaries and one of the matching Uniprot accession numbers are correctly recognized for each occurrence. The possible influence of the exact annotation schema on the performance of NER in the biomedical domain is discussed in (Alex, 2006; Shipra et al. 2004). We have also marked up mutant genes (they might be a case of free variation in the name which is unrecorded in the existing databases[7]), and special cases of ambiguity, where multiple genes on the same operon. The complete corpus is still being revised and completed, and will soon be released to the public at http://www.ebi.ac.uk/Rebholz/software.html.

6. Results

6.1 Evaluation runs

We have carried out 6 evaluation runs for the approaches introduced above on both PGN-annotated corpora. Table 3 shows the results obtained for the three methods for the human genes on the basis of the BioCreAtivE corpus.

#	Method	P	R	F
1	Direct look-up	0.03	0.82	0.06
2	DictFiltering	0.36	0.71	0.57
3	Text+DictFiltering	0.79	0.63	0.7

Table 3 Summary of evaluation results for human PGNs using the BioCreAtivE corpus.

Table 4 summarizes the results obtained for the three methods as they were applied to the identification of E. Coli PGNs.

#	Method	P	R	F
1	Direct look-up	0.14	0.45	0.21
2	DictFiltering	0.66	0.42	0.51
3	Text+DictFiltering	0.75	0.41	0.53

Table 4 Summary of evaluation results for E. coli PGNs based on the E. coli PGN corpus.

Method 1 (*Direct look-up*) involved applying the PGNs imported from the Biothesuarus directly on the corpora with two restrictions only: regular-expression based removal of corrupted and nonsense names, and only selecting PGNs relevant to either human or E. coli.

In Method 2 (*DictFiltering*) we used the dictionary filtering rules specified in section 2.2 above.

In Method 3 (*Text+DictFiltering*) we combined the dictionary filtering features with the text-driven ones.

6.2 Evaluation on the BioCreAtivE corpus

Figure 1 shows a significant increase of the F-measure despite the decrease of recall as direct matching of PGNs is restricted by dictionary filters and then context driven restrictions.

[7] Incidentally, this is where NER differs from traditional WSD in that a new sense is encountered which cannot be resolved to a pre-existing sense identifier.

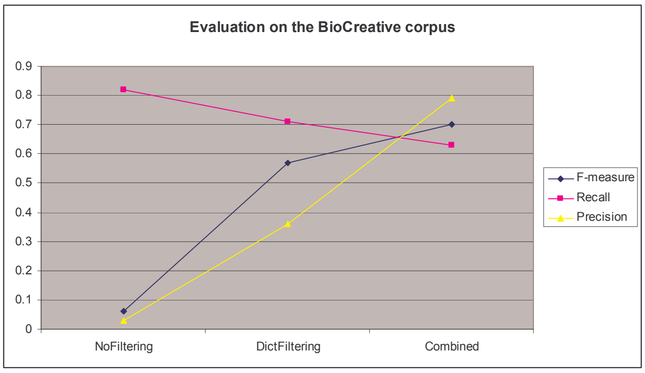
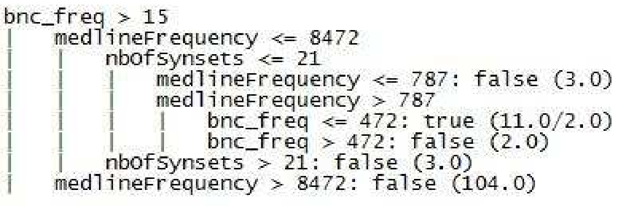


Figure 1 Precision, recall and F-measure of PGN normalization obtained based the Bio-Creative corpus.

By just applying dictionary filters based on the set of features addressing the four types of PGN polysemy introduced above, we managed to increase the precision from 0.03 to 0.36. Out of the 6 dictionary filtering features four were used by the decision tree to derive classification rules. They could be put in the following order of decreasing strength (see Table 2 for explanation of the dictionary features):

1. BNC frequency (Feature 1).
2. Medline frequency (Feature 5).
3. Number of clusters where a term occurs (Feature 2)
4. Number of distinct species taxonomy identifiers (Feature 3)

Figure 2 shows a major split in the decision space produced by the decision tree trained on static features for human PGNs.



```
bnc_freq > 15
|   medlineFrequency <= 8472
|   |   nbofSynsets <= 21
|   |   |   medlineFrequency <= 787: false (3.0)
|   |   |   medlineFrequency > 787
|   |   |   |   bnc_freq <= 472: true (11.0/2.0)
|   |   |   |   bnc_freq > 472: false (2.0)
|   |   nbofSynsets > 21: false (3.0)
|   medlineFrequency > 8472: false (104.0)
```

Figure 2 A fragment of the decision space split for human PGNs.

Table 5 shows a number of terms which have been correctly classified as false PGNs for the BioCreAtivE corpus. *Chicken* and *alternative* are common English words as indicated by their Medline and BNC frequencies. *Tissue* and *translocation* are more of biomedical terms.

Term	Polysemy type
chicken	1
alternative	1
tissue	1,4
translocation	4
p63	3
polymerase	2

Table 5 Example true negatives filtered out with static dictionary features in the gene name normalization task.

P63 is found in many species and therefore potentially polysemous in a PGN normalization task. *Polymerase* is an enzyme name which, in our Term Repository, has been assigned with as many as 87 UniProt accession numbers and 39 different species identifiers.

The outcome of the dictionary filtering phase was passed to the context-driven classifier, which further increased the precision to 0.79, while depressing the recall to 0.63. This is only natural, since in this experiment we treated the decisions made by the dictionary-based classifier as binding. In other words, a term marked as a negative in the dictionary-filtering stage could not be tagged by the context-driven classifier.

The most salient contextual features selected by the decision tree included intra-corpus ambiguity, Medline frequency, and the identification of a name as a PGN mention by Abner.

6.4 Evaluation on the E. coli corpus

Interestingly, the initial recall of E. coli PGNs is only 0.45, which may be partly due to the occurrences of mutant genes that have not been recorded in existing PGN resources used. Another major reason for the initially low recall is the occurrences of operon names, which we annotate with several identifiers matching all the genes on a given operon. As an example, we have assigned as many as 9 matching identifiers to the *TOR* (*trimethylamine N-oxide reductase*) operon. Not all of these gene names are associated with this operon in the lexical resources we have used. Yet another reason for the relatively low recall is the variability of operon names (e.g. cyoABCDE may stand for cyoA, cyoB, etc.), which occur in the corpus relatively frequently because of its gene-regulation focus. The drop in the recall as we apply the dictionary-filtering rules is rather insignificant (0.45 to 0.42) compared with the gain in the precision (from 0.15 to 0.66).

The combination of dictionary-filtering rules with the text-driven features resulted in a further increase of precision to 0.75 and a slight decrease of the recall value (0.41). The overall F-measure achieved for the E. coli PGN corpus with a combination of both methods was 0.53 (see Figure 3).

The best split of the feature space found by the decision tree for the E. coli corpus was based on a Medline frequency threshold. This criterion alone increased the precision from 0.14 to 0.66, which suggests that E. coli PGNs are much more standardized than human-relevant ones and that they do not require a complex feature set to cover specific regions of the feature space. The main split in the decision space for E. coli gene names based on Medline frequencies can be described as follows;

medlineFrequency <= 3320: true (226.0/65.0) medlineFrequency > 3320: false (65.0/14.0)

Domain-independent polysemy as indicated by the BNC frequency feature did not play an important role in the case of E. coli PGNs, since there are very few E. coli PGNs coinciding with common English words.

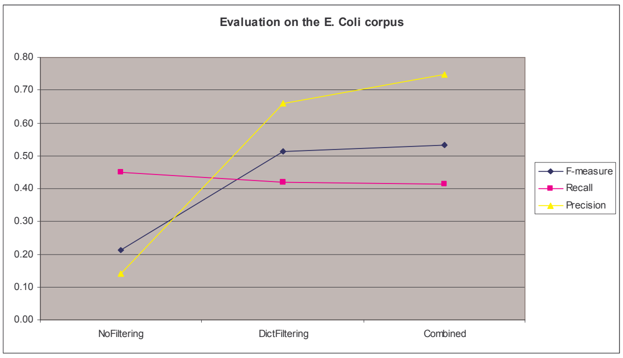


Figure 3 Precision, recall and F-measure of PGN normalization obtained based the E. coli corpus

In the decision tree generated for the context filter we find that a term’s intra-corpus ambiguity constituted the strongest feature, followed by Medline frequencies and the annotations provided by the Abner tool.

7 Conclusions

A single lexical resource integrating terms of different semantic types contains an intrinsic representation of term polysemy. Protein and gene names are the most notoriously polysemous type of named entities found in the biomedical literature and we have identified the major types of ambiguity they can introduce. We have demonstrated how a set of features that provide indications of these polysemy types can be assigned to each PGN in our Term Repository. One advantage of the set of features we propose is that they can be assigned statically to any term in a synset-based lexicon and that they are easy to map to the types of polysemy identified. In principle, the dictionary filtering features we propose can be applied to any other semantic type in our Term Repository, although so far we only provide evaluation for PGNs. Although disambiguation based on static dictionary features does not outperform fully-fledged context-driven NER, it does effectively filter out highly polysemous terms. Additionally, static dictionary features are context-independent and thus more flexibly applicable - for instance in information retrieval systems, where running intense on-the-fly disambiguation for retrieved documents could be problematic. Once they are computed and assigned for every term in the terminological resource, static polysemy indicators can be used for ad-hoc NER, conservative query expansion or relevance feedback, independently of the context in which they are retrieved. Of course, an alternative approach to polysemy-sensitive information retrieval would involve disambiguating the collection at indexing time using fully-fledged information extraction techniques.[8]

By applying dictionary filtering rules we obtained significant improvements in the F-measure of PGN identification for both corpora. As we demonstrate dictionary filtering can be further complemented with text-driven features, although the latter cannot be statically assigned to terms and need to be recomputed for the context of occurrence at hand.

Another conclusion emerges from the different precision and recall values obtained for the two corpora used for evaluation. One explanation of these differences could be that E. coli PGNs are more normalized than human PGNs (hence higher precision), but they are not as well represented in existing PGN databases (hence lower recall). Also, it seems that fewer rules are needed for the identification of truly polysemous E. coli PGNs than in the case of the more ambiguous human relevant PGNs.

8 Acknowledgements and funding

This research was sponsored by the EC STREP project “BOOTStrep” (FP6-028099, www.bootstrep.org).

Many thanks to Jung-jae Kim for his contribution to the preparation of the corpus.

9 References

Beatrice Alex, Malvina Nissim and Claire Grover, The Impact of Annotation on the Performance of Protein Tagging in Biomedical Text, In: Proceedings of LREC 2006, Genoa, Italy.

Chen L, Liu H, Friedman C., (2004), Gene name ambiguity of eukaryotic nomenclatures. *Bioinformatics*. 2005 Jan 15;21(2):248-56. Epub 2004 Aug 27.

Fellbaum Ch., (1998). *Wordnet – An Electronic Lexical Database*, MIT Press, Cambridge, Massachusetts, London England.

Hakenberg J, Bickel S, Plake C, Brefeld U, Zahn H, Faulstich L, Leser U, Scheffer T. (2005). Systematic feature evaluation for gene name recognition. *BMC Bioinformatics*. 2005;6 Suppl 1:S9.

Jimeno, A., Jimenez-Ruiz, E., Lee, V., Gaudan, S., Berlanga-Llavori, R., Rebholz-Schuhmann, R. (2008) Assessment of disease named entity recognition on a corpus of annotated sentences, BMC Bioinformatics (to appear)

Jimeno A, Pezik P., Rebholz-Schuhmann D., (2007) Information Retrieval and Information Extraction in TREC Genomics 2007. The Sixteenth Text REtrieval Conference (TREC 2007) Proceedings. NIST Special Publication: SP 500-274. Washington, USA.

Kim JJ., Pezik P., Rebholz-Schuhmann D. (2008) MedEvi: Retrieving textual evidence of relations between biomedical concepts from Medline. Bioinformatics Advance Access published online on April 9, 2008, Oxford, England.

Liu H., Hu Z., Zhang J. and Wu C., (2006). BioThesaurus:

[8] One example of a retrieval engine which combines both techniques is MedEvi (Kim et al. 2008).

a web-based thesaurus of protein and gene names. *Bioinformatics* 2006 22(1):103-105.

Morgan A., Hirschman L. (2007) Overview of BioCreAtivE II Gene Normalization. *Proceedings of the Second BioCreAtivE Challenge Evaluation Workshop*.

Quinlan, J. R. (1993) C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers

Schuemie MJ, Mons B, Weeber M, Kors JA. (2007) Evaluation of techniques for increasing recall in a dictionary approach to gene and protein name identification. *Journal of Biomedical Informatics*. 2007 Jun;40(3):316-24.

Settles B. (2005) ABNER: an open source tool for automatically tagging genes, pro-teins and other entity names in text. *Bioinformatics*. 2005 Jul 15;21(14):3191-2.

Shipra Dingare, Jenny Finkel, Christopher Manning, Malvina Nissim, and Beatrice Alex. 2004. Exploring the Boundaries: Gene and Protein Identification in Biomedical Text. Proceedings of the BioCreative Workshop, Granada.

Shipra Dingare, Jenny Finkel, Malvina Nissim, Christopher Manning and Claire Grover. 2004. A System For Identifying Named Entities in Biomedical Text: How Results From Two Evaluations Reflect on Both the System and the Evaluations. In The 2004 BioLink meeting: Linking Literature, Information and Knowledge for Biology at ISMB 2004. Republished as Shipra Dingare, Malvina Nissim, Jenny Finkel, Christopher Manning and Claire Grover. 2005. Comparative and Functional Genomics 6: 77-85.

Tanabe L, Wilbur WJ. (2002) Tagging gene and protein names in biomedical text. *Bioinformatics*. 2002 Aug;18(8):1124-32.

Wilbur J., Smith L., Tanabe L., (2007) BioCreAtivE 2. Gene Mention Task. *Proceedings of the Second BioCreAtivE Challenge Evaluation Workshop*

Pyridines, pyridine and pyridine rings: disambiguating chemical named entities

Peter Corbett[*], **Colin Batchelor**[†], **Ann Copestake**[‡]

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

Abstract

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

1. Introduction

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

2. Related Work

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

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

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

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

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

3. Annotation Guidelines

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

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

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

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

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

3.1. EXACT, CLASS and PART

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

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

3.2. SPECIES, SURFACE and POLYMER

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

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

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

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

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

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

3.3. RN, CJ and ASE

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

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

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

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

4. Inter-annotator Agreement

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

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

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

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

4.1. Results and Discussion

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

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

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

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

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

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

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

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

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

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

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

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

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

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

5. Examples

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

5.1. EXACT, CLASS and PART

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

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

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

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

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

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

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

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

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

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

Table 3: Confusion matrix for subtypes of CM.

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

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

5.2. SPECIES, SURFACE and POLYMER

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

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

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

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

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

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

6. Automated Systems

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Table 4: Automated results for CM by feature set

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

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

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

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

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

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

7. Further Directions

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

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

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

8. Conclusion

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

The annotation guidelines are available by contacting the first author.

9. Acknowledgements

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

10. References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Chemical Names: Terminological Resources and Corpora Annotation

Corinna Kolářik[*†], **Roman Klinger**[†],
Christoph M. Friedrich[†], **Martin Hofmann-Apitius**[*†], **and Juliane Fluck**[†]

[†]Fraunhofer Institute Algorithms
and Scientific Computing (SCAI)
Department of Bioinformatics
Schloß Birlinghoven
53574 Sankt Augustin, Germany

[*]Bonn-Aachen International Center
for Information Technology (B-IT)
Department of Applied Life Science Informatics
Dahlmannstrasse 2
D-53113 Bonn, Germany

corinna.kolarik@scai.fhg.de, roman.klinger@scai.fhg.de,
christoph.friedrich@scai.fhg.de, martin.hofmann-apitius@scai.fhg.de, juliane.fluck@scai.fhg.de

Abstract

Chemical compounds like small signal molecules or other biological active chemical substances are an important entity class in life science publications and patents. The recognition of these named entities relies on appropriate dictionary resources as well as on training and evaluation corpora. In this work we give an overview of publicly available chemical information resources with respect to chemical terminology. The coverage, amount of synonyms, and especially the inclusion of SMILES or InChI are considered. Normalization of different chemical names to a unique structure is only possible with these structure representations. In addition, the generation and annotation of training and testing corpora is presented. We describe a small corpus for the evaluation of dictionaries containing chemical enities as well as a training and test corpus for the recognition of IUPAC and IUPAC-like names, which cannot be fully enumerated in dictionaries. Corpora can be found on `http://www.scai.fraunhofer.de/chem-corpora.html`

1. Introduction

In life science and chemical research a huge amount of new publications, research reports and patents is produced every year. High efforts were made to improve named entity recognition (NER) to support researchers to cope with the growing amount of publications. Analysis of the quality of developed methods have been focused to a great extend on the recognition of gene and protein names. Corpora for the main model organisms have been annotated and different systems have been evaluated in international assessments.

The identification of protein and gene names is still a challenge but as a result of the mentioned efforts, dictionary and rule based methods as well as machine learning techniques are now well established for protein and gene mentions in text. The Proceedings of the BioCreative II challenge (Hirschmann et al., 2007) give a good overview about the state-of-the-art methods and their performance.

A further important entity class is composed of small chemical compounds, for instance artificial substances, like drugs, or the organism's own biomolecules like metabolites or small signaling molecules. They are analyzed in many biological, medical or pharmacological studies to clarify their effect onto biological systems or to study the biological systems on its own.

In contrast to genes coded through a nucleotide sequence and protein macromolecules coded through amino acid sequences these small chemical molecules are represented in structures. InChI and SMILES are chemical structure descriptions that have been developed to refer to a compound with a unique textual compound identifier. In addition the largest commercial chemical database (CAS) provide for its whole chemical compound content unique CAS registry numbers (e.g. 50-78-2 for Aspirin). These numbers are are often used for normalization in the chemical community but they are proprietary and contain no structural information. Because of a limited readability of such specifications for humans, trivial names or drug trade names and the nomenclature published by the *International Union of Pure and Applied Chemistry* (IUPAC, (McNaught and Wilkinson, 1997)) is commonly applied (Eller, 2006) in text. Also combinations of the different types of names as well as abbreviations, especially of often used substances, are in use.

A number of systems deal with the entity class of chemical names, spanning from manually developed sets of rules (Narayanaswamy et al., 2003; Kemp and Lynch, 1998), grammar or dictionary-based approaches (Anstein et al., 2006; Kolářik et al., 2007; Rebholz-Schuhmann et al., 2007) to machine learning based systems (Sun et al., 2007; Corbett et al., 2007).

Semantic search, classification of recognized names, or structure and substructure searches are improved by normalizing the names to the corresponding structure. Chemical dictionaries containing structural representation allows for direct mapping of recognized names to the corresponding structure at the same time. Therefore one main task during the development of dictionary based systems is the generation of comprehensive resources providing synonyms and unique identifiers for the normalization of the entities of interest.

For other representations of chemical structures like SMILES, InChI or IUPAC names such an enumeration is only possible for the most common substances. The full chemical space cannot be enumerated. Therefore dictionary independent systems are necessary for the recognition of these names. For machine learning based systems as well as for system evaluation, the annotation of text corpora is another main challenge.

To our knowledge, no general overview or evaluation on publicly available terminology resources, like databases, covering chemical entities is available. In this work, we give a survey of different data sources, and evaluate the general usability of the contained chemical terminology for Named Entity Recognition. Unfortunately, none of the corpora used for the existing approaches mentioned above is publicly available for the evaluation and development of new methods. Therefore, we annotated new corpora and provide them publicly together with the annotation guidelines on `http://www.scai.fraunhofer.de/chem-corpora.html`.

IUPAC and IUPAC-like names have been identified with a machine learning approach that is based on Conditional Random Fields (Lafferty et al., 2001). Beside trivial names, these are used most often in publications and cannot be enumerated fully in dictionaries (more details can be found in (Klinger et al., 2008)). We discuss our experiences in the generation and annotation of the corpora and give a short overview on the results.

2. Terminological Resources

Entity recognition approaches that are based on dictionaries rely on comprehensive terminology resources containing frequently used synonyms and spelling variants. An example excerpt of an extracted dictionary is given in Table 1. As for proteins and genes, databases could be a valuable resource to obtain chemical named entities and their synonyms. In this section we give an overview on available data sources.

Until recently, when the academic community started to build information sources for biologically relevant chemical compounds, chemical information was only available from commercial databases. The most important and largest resources not freely available are the CAS REGISTRY[1], the CrossFire Beilstein[2] database, and the World Drug Index[3].

For a deeper analysis we focus on freely available resources basically used in biomedicinal research. These are databases with public chemical content, thesauri and an ontology that have been growing over the last years. We concentrate on entities belonging to the class of small organic molecules and drugs from the context of human studies. Some of them contain very specific information and others cover a broad chemical space. The database PubChem[4] (Wheeler et al., 2008), the ChEBI ontology[5] (Degtyarenko et al., 2008), and MeSH[6] represent sources for a broad chemical space.

The more specialized data sources DrugBank[7](Wishart et al., 2008) and KEGG Drug[8] (Kanehisa et al., 2008) were considered as drug terminology resources. KEGG Compound[9] and the Human Metabolome Database (HMDB)[10] (Wishart et al., 2007) have been chosen as terminology resources for metabolic substances.

This survey does not claim to give a complete overview of all available chemical information resources. There is a number of other databases and resources covering specialized chemical information and a broader chemical space, e.g. UMLS[11] (Nelson et al., 2002) implying MeSH, MedlinePlus[12], and ChemIDplus[13] (Tomasulo, 2002).

2.1. Commercial Databases

CrossFire Beilstein database is a large repository for information of over 10 million organic compounds, determining their bioactivity and physical properties, ascertaining the environmental fates and their reactions. Beside structural information the entities are associated with chemical and physical facts, bioactivity data, and literature references.

CAS REGISTRY^{SM} provided by CAS, is one of the largest databases of chemical substance providing information about more than 33 million organic and inorganic substances as well as over 59 million sequences. To each substance, a unique ID (CAS Registry Number) is assigned, generated by CAS to link between the various nomenclature terms as a kind of normalization. These IDs have long been used as reference to chemicals in other databases as well as in text.

The World Drug Index contains chemical and biomedical data for over 80,000 marketed and development drugs with internationally recognized drug names, synonyms, trade names, and trivial names. Each record has a chemical structure and is classified by drug activity, mechanism of action, treatment, manufacturer, synonyms, and medical information.

2.2. Freely available Resources

From all resources introduced in this section individual dictionaries have been created and evaluated on the EVAL corpus (see Section 5.1).

PubChem consists of three linked databases – *PubChem Substance*, *PubChem Compound*, and *PubChem BioAssay*. They are part of the NCBI's Entrez information retrieval system[14]. *PubChem Compound* contains 18.4 million entries of pure and characterized chemical compounds, structure information, SMILES, InChI, and IUPAC but no further synonyms. *PubChem Substance* provides 36.8 million entries with information about mixtures, extracts, complexes, and uncharacterized substances or proteins. It comprises synonyms in the form of trivial names, brand names, IUPAC, but no SMILES, and only few mappings to InChI names. For the chemical dictionary names and synonyms as well as the chemical structure information are needed. Therefore a PubChem subset dictionary was generated with all *PubChem Substance* entries containing names, synonyms and links to corresponding entries of *PubChem Compound* (5,339,322 records).

Chemical Entities of Biological Interest (ChEBI) is a freely available controlled vocabulary of small molecular

[1] `http://www.cas.org/expertise/cascontent/registry/index.html`

[2] `http://www.beilstein.com/`

[3] `http://scientific.thomson.com/products/wdi/`

[4] `http://pubchem.ncbi.nlm.nih.gov/`

[5] `http://www.ebi.ac.uk/chebi/init.do`

[6] `http://www.nlm.nih.gov/mesh/meshhome.html`

[7] `http://drugbank.ca/`

[8] `http://www.genome.jp/kegg/drug`

[9] `http://www.genome.jp/kegg/compound`

[10] `http://hmdb.ca/`

[11] `http://www.nlm.nih.gov/research/umls/`

[12] `http://medlineplus.gov/`

[13] `http://chem.sis.nlm.nih.gov/chemidplus/`

[14] `http://www.ncbi.nlm.nih.gov/`

i	id_i	S_i
1	DB06151	CC(=O)NC(CS)C(=O)O; InChI=1/C5H9NO3S/c1-3(7)6-4(2-10)5(8)9/h4,10H,2H2,1H3,(H,6,7)(H,8,9)/t4-/m0/s1/f/h6,8H; Acetylcysteine; ACC; Mucomyst; Acetadote; Fluimucil; Parvolex; Lysox; Mucolysin; (2R)-2-acetamido-3-sulfanylpropanoic acid; ...
2	DB05246	CC1(CC(=O)N(C1=O)C)C2=CC=CC=C2; InChI=1/C12H13NO2/c1-12(9-6-4-3-5-7-9)8-10(14)13(2)11(12)15-/h3-7H,8H2,1-2H3; Methsuximide; Petinutin; Celontin; 1,3-dimethyl-3-phenylpyrrolidine-2,5-dione; ...

Table 1: Example for a dictionary based on DrugBank, usually incorporated in rule based Named Entity Recognition systems. The identifier (in this case a DrugBank identifier) is denoted with id_i, the set of synonyms with S_i.

Resource	Number of entries
CrossFire Beilstein	10 mio.
CAS	33 mio.
World Drug Index	80,000
PubChem_C; PubChem_S	18.4 mio.; 36.8 mio.
MeSH_T	8,612
MeSH_C	175,136
ChEBI	15,562
KEGG (K-C; K-D)	21,498 (15,033; 6,834)
DrugBank	4,764
HMDB	2,968

Table 2: Total number of entities contained in chemical information resources (PubChem_C: PubChem Compound; PubChem_S: PubChem Substance; K-C: KEGG-compound; KEGG-Drug)

entities that intervene in the processes of living. Entities are organized in an ontological classification and are grouped by their chemical structure and functional properties. General chemical class terms, biological and pharmacological functions, and compounds with general names are covered as well as synonyms of the form of trivial name, IUPAC, and sum formula. For most of the chemical compounds SMILES and InChI names are given. We used the release version 35 of ChEBI provided in the OBO-format.

Medical Subject Headings (MeSH) is a controlled vocabulary thesaurus from the National Library of Medicine (NLM)[15]. It is used by NLM for indexing articles from the MEDLINE PubMED database as well as a catalog database for other media of the library. The terms are organized in a hierarchy to which synonyms as well as inflectional term variants are assigned. A subset of the MeSH thesaurus (version 2007 MeSH) covering the chemical category of MeSH (tree concepts with node identifiers starting with 'D') was extracted to give one dictionary of MeSH (referenced further as MeSH_T). Furthermore, NLM provides a compound list with over 175,000 entries containing synonyms like trivial and brand names, IUPAC and abbreviations which was used to generate another dictionary, referenced further as MeSH_C.

Kyoto Encyclopedia of Genes and Genomes (KEGG) is a composite database that integrates genomic, chemical, and systemic functional information. Two sub-databases – *KEGG COMPOUND* and *KEGG DRUG* – are considered to be terminology resources for the dictionary creation. The types of compounds provided by *KEGG COMPOUND* span from single ions (e.g. Mg^{2+}), simple compounds (like different sugars or cofactors of enzymes, metabolites, products of microorganisms, or nuclear receptor compounds like GW 6471) to peptides and basic RNAs – all essential endogenous molecules of cells. *KEGG DRUG* covers all approved drugs in the United States of America and Japan. Every entry of both databases is linked to a unique chemical structure and to standard generic names that could be of the type IUPAC and trivial name. For the creation of the two dictionaries KEGG_C and KEGG_D the fields 'NAME', 'FORMULA', and 'DBLINKS' of the KEGG proprietary format files *compound* and *drug* have been used.

DrugBank is a specific database about pharmaceuticals, that combines detailed chemical, pharmacological, and pharmaceutical information with drug target information. It provides trivial, brand, and brand mixture names, IUPAC and a structure for almost every entity as SMILES or InChI. DrugBank is available as a single file in a proprietary format. Following fields have been extracted: 'Name', 'Synonyms', 'Brand Names', 'Brand Mixtures', 'Chemical IUPAC Name', 'Chemical Formula', 'InChI Identifier', 'Isomeric SMILES', 'Canonical SMILES', and 'CAS Registry Number'.

Human Metabolome Database (HMDB) is a freely available database containing detailed information about small molecule metabolites found in the human body. The focus lies on quantitative, analytic or molecular scale information about metabolites, their associated enzymes or transporters and their disease-related properties. The database currently contains nearly 3000 metabolite entries, like hormones, disease-associated metabolites, essential nutrients, and signaling molecules as well as ubiquitous food additives and some common drugs. HMDB is downloadable as a single file with a similar proprietary format as DrugBank. Following fields have been extracted: 'Name', 'Common Name', 'Synonyms', 'Chemical IUPAC Name', 'Isomeric SMILES', 'Canonical SMILES', 'InChI Identifier', and 'CAS Registry Number'.

3. Analysis of the Chemical Information Resources

In this section we discuss the general usability of the above mentioned resources for dictionary based named entity recognition approaches. The resources were analyzed with

[15] `http://www.nlm.nih.gov/`

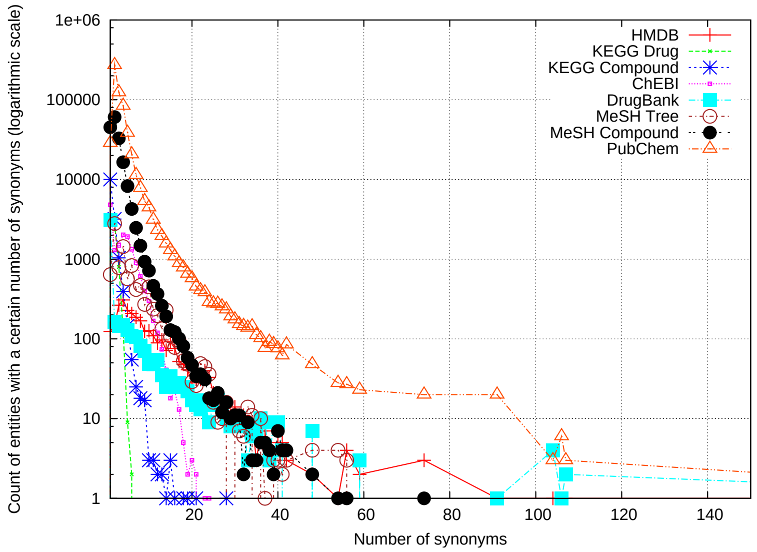


Figure 1: Plot of the synonym count distribution for the analyzed databases

	PubChem subset	MeSH_T	ChEBI	DrugBank	HMDB	MeSH_C	KEGG
SMILES	4,080,909	—	8,371	4,489	2,881	—	—
InChI	4,080,909	—	8,280	4,486	2,859	—	17,021
CAS	397,858	—	4,566	2,223	2,527	175,136	13,545
Percentage of synonyms covered by PubChem	100 %	22 %	56 %	66 %	54 %	28 %	79 %
Cross references	yes	—	yes	yes	yes	—	yes
Total No. of entries	5,339,322	8,612	15,562	4,764	2,968	175,136	21,498

Table 3: Overview of the linkage of the entities to structure information in the analyzed data sources. For PubChem only *PubChem Substance* entries containing a *PubChem Compound* link were included. For KEGG the respective values of the drug and compound sub-databases were unified. For the PubChem coverage all synonyms of all entries are compared.

regard to following properties:

- Total number of entries,
- Provided number of synonyms,
- Linkage to a structure, and
- Cross linkage to other databases.

Table 2 gives an overview about the total amount of the entities provided by the analyzed sources. All commercial databases contain a huge number of chemical entities, reflecting their growth for a long time. In comparison to them, PubChem is the biggest collection of public chemical data, followed by MeSH compounds. The remaining specialized resources, like DrugBank or HMDB, contain fewer entities but highly comprehensive biomedical information about them. Figure 1 reflects the distribution of the occurrences of synonyms for every analyzed resource. Most entries contain only few synonyms. Entries of PubChem, both MeSH dictionaries, and DrugBank as well as HMDB contain a high amount of synonyms. A high number of provided synonyms is of high value for the creation of the dictionaries. A comprehensive coverage of the chemical terms and their synonyms used in text leads to a good performance of a dictionary-based NER approach by avoiding a high false negative rate. Comparison of the synonyms contained in PubChem to the other databases (cf. Table 3) showed that there are differences in the synonym coverage in the analyzed resources. About 79 % of the KEGG synonyms are included in PubChem and 55 % of the CheBI entities but only 22 % of the MeSH tree synonyms could be found. Combining all analyzed dictionaries, 69 % of the synonyms are not from PubChem but from the other resources. Hence, it is meaningful to use an all-integrating dictionary. instead of incorporating only PubChem.

Table 3 presents the number of the resource which are mapped to InChI, SMILES or CAS. Unique representations are relevant for the mapping and normalization of the identified chemical names from text to a chemical structure. All entries in the selected PubChem subset contain InChI information and two third of the entries contain the CAS registry number. Most entries in DrugBank, HMDB, ChEBI are mapped to all three chemical representations and in KEGG, InChI and CAS registry numbers are included. All entries of MeSH_C are mapped to CAS identifiers but no other chemical representations like InChI is given.

In addition no cross references to other data sources are included. The other sources contain a high number of cross references and references to PubChem, KEGG and ChEBI are given in all databases. PubChem contains the highest number of cross references and in addition links to MeSH.

4. Annotation and Corpus Generation

For evaluation purposes of NER-systems as well as for the training of machine learning based methods annotated corpora are needed. Corbett et al. (2007) describe a corpus annotation, but the corpus as well as the annotation guidelines are not publicly available. Because annotated corpora for the chemical domain are not public available yet, we describe three corpora consisting of MEDLINE abstracts. A small evaluation corpus (EVAL corpus) containing entities of all classes described in Tables 3 and 4.1 has been annotated to give an overview of the different chemical name classes found in MEDLINE text. This corpus will be used for a first assessments of chemical dictionaries and for the evaluation of methods for chemical name recognition. In addition, a training and a test corpus was generated for the machine learning based recognition of IUPAC and IUPAC-like names and has been annotated with the classes IUPAC and PART. In the following sections our assignment of chemical terms to various defined annotation classes and the corpus annotation is described.

4.1. Chemical Entity Classes used for the Annotation

To allow an annotation even for non-chemical experts a simplified classification schema with respect to chemical classification was developed. The defined classes are IUPAC, PART, TRIVIAL, ABB, SUM, and FAMILY, shown in Table 3 with descriptions and examples. The separation between TRIVIAL and IUPAC names is based on the term length, names with only one word were classified as TRIVIAL even if they were IUPAC names. Multi word systematic and semi-systematic names are always annotated as IUPAC. This includes names that imply only a IUPAC-like part (e.g. 17-alpha-E) or names including a labeling (e.g. 3H-testosterone). This does not follow strictly the definition of IUPAC, but such terms are less likely contained in databases and cannot be found with a pure dictionary-based approach. For the correct resolution of enumerations, partial chemical names have been annotated separately as PART, but chemical names were not tagged in other entities (e.g. in protein names). Names were only tagged as FAMILY if they describe well defined chemical families but not pharmacological families (e.g. glucocorticoid was labeled but not anti-inflammatory drug). Substances used as base for building various derivates and analogs were tagged as IUPAC, not as FAMILY (e.g. 1,4-dihydronaphthoquinones). More examples and their labels used for the annotation are provided for clarification in Table 4.1. All defined classes were used for the annotation of the evaluation corpus. This annotation allows the assessment of distribution of chemical names in MEDLINE text and the coverage of the different dictionaries and recognition approaches. We do not imply to use this classification as final annotation scheme for chemical name annotation. Further iterations of evaluation and annotation are necessary and are work in progress including more chemical experts.

4.2. Corpus Selection for the Annotation and Evaluation of all Chemical Classes

Based on the assumption that abstracts containing IUPAC names also contain other nomenclatures, a preliminary system for detecting IUPAC names as described in Section 4.3 (Klinger et al., 2008) was applied to select abstracts from MEDLINE containing at least one found entity. Next to abstracts selected with this procedure, we selected abstracts containing problematical cases as well as those containing no entities. This procedure formed a corpus of 100 abstracts containing 391 IUPAC, 92 PART, 414 TRIVIAL, 161 ABB, 49 SUM, and 99 FAMILY entities.

4.3. Corpus Generation for the Recognition of IUPAC and IUPAC-like Entities

As a training corpus for a Conditional Random Field (CRF), 463 abstracts have been selected from 10,000 sampled abstracts from MEDLINE. It was annotated by two independent annotators. A conclusive training corpus was generated using a combination of both annotations by an independent person. This resulted in a corpus containing 161,591 tokens with 3,712 IUPAC annotations. Here, the class PART was included in the class IUPAC due to morphological similarity of these classes which is important for the machine learning approach described in Section 5.2.

A test corpus was selected to test the system trained on the above described training corpus. For that, 1000 MEDLINE records with 124,122 tokens were sampled equally distributed from full MEDLINE and has been annotated. It comprises 151 IUPAC entities. The sampling process ensures to have representative text examples of the full MEDLINE. This is especially beneficial for a correct analysis of the false positives.

4.4. Inter-Annotator Agreement

For the corpus with all chemical entities described in Section 4.2 and the training corpus described in Section 4.3, the inter-annotator agreement was evaluated.

Recognizing the boundaries without considering the different classes on the test corpus described in Section 4.2, the inter-annotator F_1 is 80 % and for the IUPAC entity in the training corpus, the F_1 measure is 78 %. For both corpora conclusive corpora were generated. The conclusive training corpus and the first-annotated corpus differ to a lower degree, the inter-annotator F_1 measure is 94 %. In contrast Corbett et al. (2007) claimed 93 % for the training corpus for the system *OSCAR*. One reason for the lower F_1 measure in the first annotation in comparison to the result of Corbett and his colleagues is our differentiation of the IUPAC entity to other chemical mentions. The appropriate usage of those is not always easy to decide, while all chemical mentions in the corpus generated by Corbett are combined in one entity (see Section 5.2 for more details). Another reason is the different experience level of our annotators. One annotator participated in the development of the annotation rules. The corpus was annotated partly by this person more than once during this process. The second person annotated the whole set based on the annotation guideline without an intermediate revision. Therefore, we propose a two step process

Chemical Class	Description	Example Annotation
IUPAC	IUPAC names, IUPAC-like names, systematic, and semi-systematic names	1-hexoxy-4-methyl-hexane
PART	partial IUPAC class names	17beta-
TRIVIAL	trivial names	aspirin, estragon
ABB	abbreviations and acronyms	TPA
SUM	sum formula, atoms, and molecules, SMILES, InChI	KOH
FAMILY	chemical family names	disaccaride

Table 4: Chemical entity classes used for the corpora annotation

Name	Labeled Sequence	Label	Explanation
Acetylsalicylate	Acetylsalicylate	TRIVIAL	
elaidic acid	elaidic acid	IUPAC	multi word systematic and semi systematic names are labeled as IUPAC
testosterone	testosterone	TRIVIAL	
3H-testosterone	3H-testosterone	IUPAC	contains part IUPAC-like structure (3H-);
17-alpha-E	17-alpha-E	IUPAC	E = chemical abbreviation
17beta-HSD	—	—	HSD = protein name
N-substituted-pyridino[2,3-f] indole-4,9-dione	N-substituted-pyridino[2,3-f] indole-4,9-dion	IUPAC	
2-acetyloxybenzoic acid	2-acetyloxybenzoic acid	IUPAC	
Ethyl O-acetylsalicylate	Ethyl O-acetylsalicylate	IUPAC	
pyrimidine	pyrimidine	FAMILY	
1,4-dihydronaphthoquinones	1,4-dihydronaphthoquinones	IUPAC	
Ca(2+)	Ca(2+)	SUM	
(14)C	(14)C	SUM	

Table 5: Annotation examples

for further annotations. In a first step an inter-annotator agreement should be build only on a small set of annotated abstracts and discrepancies could be reviewed with all annotators. Then the larger set of abstracts could be annotated with higher confidence.

5. Recognition of Chemical Names

5.1. Dictionary-based Recognition of Chemical Compounds

Dictionaries built from the different terminological resources were used to recognize chemical entities in the EVAL corpus. Following constraints were used for all searches:

- No curation of the created raw dictionaries was done, which means that no names were removed, added or changed.
- All synonyms were searched with a simple case insensitive string search, dashes were ignored.
- No control of the correct association of the found names to the corresponding entry was performed.

The results with uncurated dictionaries and such a simple search strategy should only give a rough estimate of the coverage of different sources and the efforts which have to be invested in curation and search strategies.

The search results obtained with every individual dictionary and a combination of the results of all dictionaries are provided in Table 6. The first two rows show precision and recall on a combination of all annotation classes. The rates in brackets were obtained when also partial matches were considered as true positives. The highest precision rates were achieved by the KEGG Drug dictionary (59 %) followed by the MeSH_C dictionary (44 %). The lowest precision of 13 % and 15 % was obtained by ChEBI and PubChem respectively. Many unspecific terms are contained in ChEBI (e.g. groups or inhibitors), and also terms that have not been annotated as a chemical family term (e.g. enzyme inhibitors or adrenergic agonist). Such terms were considered to be pharmaceutical property terms. Additionally, many other names are unspecific, like one-character tokens (e.g. D, J) and common word names (e.g. at, all). Therefore, we conclude that curation processes are necessary to achieve a higher performance with the dictionaries. Experiences with the gene and protein name recognition (Hirschmann et al., 2007) let us assume that the precision could be highly enhanced through dictionary curation and more elaborate named entity recognition techniques.

The recall of the dictionary based named entity recognition is low. The highest recall was obtained with the PubChem dictionary identifying 33 % of all entries, followed by the ChEBI and the MeSH_T dictionary (both 27 %). The conflation of all search results enhances the recall to 49 %, but decreases the precision to 13 %. The participation of the different dictionaries on the combined result has to be checked further for recall and precision.

Class	PubChem	ChEBI	MeSH_C	MeSH_T	HMDB	KEGG_C	KEGG_D	DrugBank	Combined
ALL (1206)	*0.15 (0.26)* 0.33 (0.60)	*0.13 (0.34)* 0.27 (0.68)	*0.44 (0.64)* 0.10 (0.15)	*0.34 (0.42)* 0.27 (0.34)	*0.21 (0.44)* 0.16 (0.33)	*0.30 (0.54)* 0.24 (0.43)	*0.59 (0.76)* 0.12 (0.16)	*0.33 (0.43)* 0.13 (0.17)	*0.13 (0.22)* 0.49 (0.85)
IUPAC (391)	0.16 (0.69)	0.08 (0.85)	0.09 (0.21)	0.05 (0.29)	0.06 (0.44)	0.07 (0.51)	0.03 (0.17)	0.01 (0.17)	0.23 (0.94)
PART (92)	0.04 (0.32)	0.13 (0.72)	0.00 (0.05)	0.00 (0.01)	0.04 (0.32)	0.05 (0.24)	0.00 (0.00)	0.00 (0.00)	0.13 (0.75)
SUM (49)	0.31 (0.73)	0.31 (0.88)	0.04 (0.08)	0.00 (0.00)	0.00 (0.30)	0.12 (0.46)	0.00 (0.00)	0.00 (0.00)	0.31 (0.88)
TRIV (414)	0.66 (0.82)	0.52 (0.78)	0.18 (0.19)	0.64 (0.65)	0.36 (0.42)	0.57 (0.64)	0.35 (0.36)	0.40 (0.41)	0.88 (0.97)
ABB (161)	0.49 (0.72)	0.23 (0.55)	0.09 (0.11)	0.2 (0.23)	0.15 (0.34)	0.15 (0.32)	0.03 (0.03)	0.03 (0.03)	0.58 (0.83)
FAM (99)	0.18 (0.5)	0.42 (0.69)	0.05 (0.09)	0.42 (0.42)	0.08 (0.13)	0.19 (0.35)	0.17 (0.03)	0.00 (0.03)	0.71 (0.89)

Table 6: Comparison of the entities found in the evaluation corpus with dictionaries based on the analyzed resources. All annotation classes are considered. (The total number of the annotated entities per class are given in brackets.) Precision (slanted) and recall are given for an exact match of an entity and a match where the identification of a subset of the term is sufficient (values behind the recall values in brackets).

The analysis of the recall for every single annotation class confirms our hypothesis that names belonging to the Trivial class could be found with the highest recall. The search with the PubChem dictionary identified 66 %, followed by MeSH_T with 64 % and KEGG_C with 57 %. The combination of the results lead to a promising recall of 88 %. Considering the recognition of family names by the ChEBI and the MeSH_T dictionary obtained the highest value (both 42 %). This is not very remarkable, because only those two resources contain general chemical group and family terms in their hierarchy. Sum formulas (mainly annotated as shown in Table 4.1) were only recognized to a certain degree by ChEBI, PubChem (both 31 %), and KEGG_C dictionary (12 %). The recognition rate of the Abb class has to be taken with caution because abbreviations are often short names, sometimes only one character long and therefore highly ambiguous.

As we previously assumed, Iupac names have been recognized with a low recall by all tested dictionaries. The partial match rate is high, especially for the PubChem and ChEBI dictionary. Some partial matches, e.g. 'testosterone' in '3H-testosterone', could be accepted, but many terms, e.g. diethyl or benzoyl being part of 'diethyl N-[2-fluoro-4-(prop-2-ynylamino)benzoyl]-L-glutamate', increase the rate of false positive partial matches. Therefore, strategies need to be integrated for an efficient recognition system to avoid such problems.

In summary we can conclude from this experiment that the recall of a simple search strategy that uses the individual uncured dictionaries is low. The combination of all dictionaries leads to an acceptable rate for Trivial and Family names but not for Iupac and Part names. For the recognition of the latter two a machine learning approach might be advantageous compared to a dictionary approach. Thus a machine learning based strategy for the Iupac name recognition is described in the next section.

	Precision	Recall
IUPAC tagger on test corpus sampled from MEDLINE (IUPAC + PART entities)	86.50	84.80
IUPAC tagger on Eval corpus (all entities)	91.41	29.04
IUPAC tagger on Eval corpus (IUPAC + PART entities)	81.38	73.18
IUPAC tagger on Eval corpus (IUPAC entities)	—	77.11
IUPAC tagger on Eval corpus (PART entities)	—	41.18
IUPAC tagger on Eval corpus (TRIVIAL entities)	—	8.42
OSCAR on Eval corpus with all entities	52.09	71.86

Table 7: Results of the machine learning-based tagger and of the system *OSCAR* for IUPAC entities and all entities on the Eval corpus and on the test corpus sampled from MEDLINE (in %).

5.2. CRF-based Iupac Name Recognition

To improve the recognition of Iupac names, the training corpus described in Section 4.3 was used to train a Conditional Random Field. Due to the morphological similarity of Iupac and Part entities they have been combined leading to a system that does not separate these two classes. The parameter optimization (e.g. feature selection) is described in detail in Klinger et al. (2008).

An evaluation on the sampled test corpus of 1000 abstracts from MEDLINE shows an F_1 measure of 85.6 % with a precision of 86.5 % and a recall of 84.8 %. Applying this tagger on the EVAL corpus with several entity classes described in Section 4.2, it recognizes 73.18 % of the IUPAC and IUPAC-like names with a precision of 81.38 % (considering only IUPAC and PART names as true positive hits). The recall on the separated classes IUPAC and PART (namely 77.11 % and 41.18 %) on the EVAL corpus motivates the combination of these classes for machine learning purposes.

The precision of 91.41 % on all entities is much higher than only on the IUPAC entities due to the recognition of 8.42 % of the TRIVIAL class names. They are frequently used within IUPAC terms and cannot be easily separated by the system. A separation from the other classes ABB, SUM, and FAMILY is perfectly given.

It needs to be analyzed if trivial names could be recognized with a machine learning based method with similar performance to enhance the recall of the system which is now at 29 % considering all chemical classes. Here, an additional annotation of the training set is necessary.

To compare the *OSCAR* software, this approach was also used for the recognition of all entities in the EVAL corpus. *OSCAR* has an overall high recall of almost 72 % accompanied with a precision of 52 %. The recall is similar to the reports in (Corbett et al., 2007) (73.5 % recall, 75.3 % precision) but the precision is lower. We did not analyze the results in detail but certainly one reason for the lower precision can be found in the different annotation of chemical entities underlying the training corpus used in *OSCAR*. One difference is for example the annotation of more general annotation of chemical names (e.g. dry ice).

6. Conclusion

To a certain amount, trivial names and family names but not IUPAC like names are covered by the different chemical resources analyzed in this paper. PubChem, as the largest resource, does not include all names covered by the smaller sources. Hence, the combination of the search results from all terminologies lead to a high increase in recall, especially for family and trivial names. The development of a training corpus for IUPAC like entities lead to a performant CRF-based IUPAC tagger.

These results are motivating for further investigations in the generation of dictionaries as well as testing different annotation classes to be used for training and the combination of machine learning based chemical name recognition and dictionary based normalization of chemical names.

7. Acknowledgments

We would like to thank Theo Mevissen for dictionary generation and technical support and Harsha Gurulingappa and Erfan Younesi for corpus annotation. The work of Roman Klinger is funded by the MPG-FhG Machine Learning Collaboration (`http://lip.fml.tuebingen.mpg.de/`).

8. References

S. Anstein, G. Kremer, and U. Reyle. 2006. Identifying and classifying terms in the life sciences: The case of chemical terminology. In Nicoletta Calzolari, Khalid Choukri, Aldo Gangemi, Bene Maegaard, Joseph Mariani, Jan Odijk, and Dnaiel Tapias, editors, *Proc. of the Fifth Language Resources and Evaluation Conference*, pages 1095–1098, Genoa. Italy.

P. Corbett, C. Batchelor, and S. Teufel. 2007. Annotation of chemical named entities. In *BioNLP 2007: Biological, translational, and clinical language processing*, pages 57–64, Prague, June.

K. Degtyarenko, P. de Matos, M. Ennis, J. Hastings, M. Zbinden, A. McNaught, R. Alcántara, M. Darsow, M. Guedj, and M. Ashburner. 2008. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, 36(Database issue):D344–D350, Jan.

G. A. Eller. 2006. Improving the quality of published chemical names with nomenclature software. *Molecules*, 11:915–928.

L. Hirschmann, M. Krallinger, and A. Valencia, editors. 2007. *Proc. of the Second BioCreative Challenge Evaluation Workshop*. Centro Nacional de Investigaciones Oncologicas, CNIO.

M. Kanehisa, M. Araki, S. Goto, M. Hattori, et al. 2008. Kegg for linking genomes to life and the environment. *Nucleic Acids Res*, 36(Database issue):D480–D484, Jan.

N. Kemp and M. Lynch. 1998. The extraction of information from the text of chemical patents. 1. identification of specific chemical names. *Journal of Chemical Information and Computer Sciences*, 38(4):544–551.

R. Klinger, C. Kolářik, J. Fluck, M. Hofmann-Apitius, and C. M. Friedrich. 2008. Detection of IUPAC and IUPAC-like Chemical Names. *Bioinformatics*. Proceedings of the International Conference Intelligent Systems for Molecular Biology (ISMB). accepted.

C. Kolářik, M. Hofmann-Apitius, M. Zimmermann, and J. Fluck. 2007. Identification of new drug classification terms in textual resources. *Bioinformatics*, 23(13):i264–i272.

J. D. Lafferty, A. McCallum, and F. C. N. Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proc. of the Eighteenth International Conference on Machine Learning (ICML 2001)*, pages 282–289. Morgan Kaufmann Publishers.

A. D. McNaught and A. Wilkinson. 1997. *Compendium of Chemical Terminology – The Gold Book*. Blackwell Science.

M. Narayanaswamy, K. E. Ravikumar, and K. Vijay-Shanker. 2003. A biological named entity recognizer. In *Proc. of the Pacific Symposium on Biocomputing*, pages 427–438.

S. J. Nelson, T. Powell, and B. L. Humphreys. 2002. The unified medical language system (umls) project. In Allen Kent and Carolyn M. Hall, editors, *Encyclopedia of Library and Information Science*, pages 369–378. Marcel Dekker, Inc, New York.

D. Rebholz-Schuhmann, H. Kirsch, M. Arregui, S. Gaudan, M. Riethoven, and P. Stoehr. 2007. EBIMed – text crunching to gather facts for proteins from medline. *Bioinformatics*, 23:237–244.

B. Sun, Q. Tan, P. Mitra, and C. L. Giles. 2007. Extraction and search of chemical formulae in text documents on the web. In *Proc. of the International World Wide Web Conference*, pages 251–260, May.

P. Tomasulo. 2002. ChemIDplus – super source for chemical and drug information. *Med Ref Serv Q*, 21(1):53–59.

D. L. Wheeler, T. Barrett, D. A. Benson, et al. 2008. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 36:D13 – D21, January.

D. S. Wishart, D. Tzur, C. Knox, R. Eisner, et al. 2007. HMDB: the human metabolome database. *Nucleic Acids Res*, 35(Database issue):D521–D526, Jan.

D. S. Wishart, C. Knox, A. C. Guo, et al. 2008. DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res*, 36(Database issue):D901–D906, Jan.

Towards a Human Anatomy Data Set for Query Pattern Mining based on Wikipedia and Domain Semantic Resources

Pinar Oezden Wennerberg [1, 2], Paul Buitelaar [3], Sonja Zillner[1]

[1]Siemens AG, Corporate Technology, Knowledge Management CT IC 1
Otto-Hahn-Ring 6, 81739, Munich
Germany

[2]Ubiquitous Knowledge Processing Lab TU Darmstadt
FB 20, Hochschulstraße 10, D-64289 Darmstadt,
Germany

[3]DFKI GmbH. Language Technology Lab
Stuhlsatzenhausweg 3 D-66123, Saarbrücken,
Germany

pinar.wennerberg.ext@siemens.com, paulb@dfki.de, sonja.zillner@siemens.com

Abstract

Human anatomy knowledge is an integral part of radiological information, which is necessary for image annotation in a semantic cross-modal image and information retrieval scenario. Anatomy and radiology related concepts and relations can be discovered from an anatomy corpus, which can be build up from Wikipedia as reported here. An ontology of human anatomy and a controlled vocabulary for radiology are used as knowledge resources in the search of significant concepts and relations. Our ultimate goal is to use the concepts and the relationships discovered in this way to identify potential query patterns. These query patterns are the abstractions of the actual queries that radiologists and clinicians would typically pose to a semantic search engine to find patient-specific sets of relevant images and textual data.

Introduction

This paper describes ongoing work towards the development and use of a human anatomy data set based on Wikipedia and domain semantic resources in human anatomy and radiology. In particular, we are using the Foundational Model of Anatomy[1] or 'FMA' (Rosse and Mejino, 2003) and the Radiology Lexicon[2] or 'RadLex' (Langlotz, 2006) for this purpose.

Ultimately we aim at using the data set that we are constructing for the derivation of query patterns that would typically be used by clinicians and radiologists to find patient-specific sets of relevant images, i.e. images that show similar conditions and/or symptoms. The context of our work is in the Theseus-MEDICO[3] project, which has a focus on cross-modal image and information retrieval in the medical domain.

The focus of the work reported here is on setting up a Wikipedia-based corpus of human anatomy texts and the statistical profiling of FMA (human anatomy) and RadLex (radiology) terms on the basis of this resource. Using this information we will then be able to extract relations that are likely to occur between statistically relevant terms - and the concepts they express. The final goal of our work will be to derive potential query patterns from the extracted set of relations that can be used in the MEDICO semantic-based image retrieval application.

The remainder of this paper is organized as follows. The first section outlines the context of our work in the MEDICO project. In section 2 we compare some related work, while in the third section we describe the two domain specific semantic resources used, i.e. the FMA and RadLex. Here we also describe the human anatomy corpus that we derived from Wikipedia and the steps taken to construct it. Section 4 provides details on the statistical profiling of the FMA and RadLex terms on basis of the corpus, including the correspondence of terms from both resources. The final section includes the conclusions and our plans for future work.

1. MEDICO

MEDICO - Towards Scalable Semantic Image Semantics - is an application scenario of the THESEUS Program funded by the German Federal Ministry of Economics and Technology. MEDICO addresses the need for advanced image searching technologies enabling the direct access to and seamless integration of image semantics. Through the rapid advances in imaging technologies, more and more medical image data is generated by hospitals, pharmaceutical companies, and medical research.

There exist a wide range of different imaging technologies and modalities, such as 4D 64-slice Computer Tomography (CT), whole-body Magnet Resonance Imaging (MRI), 4D Ultrasound, and the fusion of Positron Emission Tomography and CT (PET/CT) providing detailed insight into human anatomy, function, and disease associations. Moreover, advanced techniques for analyzing imaging data generating additional

[1] http://sig.biostr.washington.edu/projects/fm/
[2] http://www.rsna.org/radlex/
[3] http://theseus-programm.de/scenarios/en/medico

quantitative parameters pave the way for improved clinical practice and diagnose. However, for advanced applications in Clinical Decision Support and Computer Aided Diagnoses the comparative exploration of similar patient information is required. The missing link here fore is a flexible and generic image understanding. Currently, the large amounts of heterogeneous image data are stored in distributed and autonomous image databases being indexed by keywords without capturing any semantics.

The objective of MEDICO is to build the next generation of intelligent, scalable and robust search engine for the medical imaging domain. By integrating higher level knowledge represented by ontologies, the different semantic views of the same medical images, such as structural aspects, functional aspects, and disease aspects, can be reflected and integrated. Combining semantics with image understanding facilitates the formal description of bridges between different domains that can be used for comparative patient data exploration. The overall goal of MEDICO is to empower the imaging content-stakeholders, i.e. clinicians, pharmaceutical specialists, patient citizens, and policy makers, by providing flexible and scalable semantic access to medical image databases.

For a beneficial integration of external semantics within advanced image search, one has to decide which external knowledge resources are appropriate for the purpose in mind, i.e. which external knowledge source captures the relevant knowledge in an appropriate level of detail for a particular context. Within the MEDICO project, one of the selected use case scenarios aims for improved image search in the context of patients suffering of lymphoma in the neck area. Lymphoma, a type of cancer originating in lymphocytes, is a systematic disease with manifestations in multiple organs.

As imaging is done several times through the course of disease and different imaging modalities are used, scalable and flexible image search for lymphoma is of particular relevance: scalable image search functionalities can be easily extended to other body regions and flexibility achieved by incorporating semantics allows to use all the heterogeneous patient information, such as imaging features, symptoms, or lab results, for diagnosis and prediction of disease development.

To enable improved image search and advanced medical applications, it is relevant to find out what kind of knowledge the clinician, e.g. a lymphoma expert, wants to know or the queries that clinicians are interested in. Existing methodologies for knowledge engineering, such as (Schreiber et al., 2000) and (Sure et al., 2002), following systematic interview-based approaches for the analysis of knowledge requirements were not suitable for our task in mind. As clinicians speak (and think) a very different language than computer scientists and as they are always in lack of time, the analysis of knowledge requirements by interviews were not feasible.

Moreover, as we are aiming to develop next generation image search facilities, there was the danger that our addressed users are too constrained in their imagination by familiar, existing systems. For example, it is not possible for the current pattern recognition algorithms to produce image annotations that express the *Lymph node* as *located_in* the *Neck* that *has_dimension X*. In real life however, clinicians and radiologists look for information and images that report "*an enlargement in the dimension of the lymph node in the neck*", which is an essential radiological pattern to re-stage a certain type of lymphoma in the head and neck region. Therefore, within our approach, we first establish, semi-automatically, hypotheses about possible user queries that are the so called query patterns. These patterns are derived from a combination of certain constraints and joint view of anatomy and radiology. More concretely, the joint view corresponds to the anatomical structures found on radiology images e.g. the CT scan of the neck region. The constraints are e.g. the spatial relations such as *located_in* that constrain the model by asserting that certain anatomical structures are (only) expected to be in certain other ones. For example, only a certain kind of lymph node will be found in the neck region. Accordingly, an example query pattern might look like this:

[ANATOMICAL STRUCTURE]	*located_in*	[ANATOMICAL STRUCTURE]
	AND	
[ANATOMICAL STRUCTURE]	*has_dimension*	[DIMENSION]

Once an initial set of similar patterns has been established in this way, they will be evaluated by clinicians for their validity and relevance.

2. Related work

As far as we know the data set that we are developing will be the first one based on a text corpus specifically on human anatomy, in particular from the viewpoint of radiology. Anatomy data sets that do exist are image-centered and consist of collections of radiology images without associated textual data, such as the Visible Human Project data set[4]. A future outcome of our research could in fact be the compilation of a multimedia data set consisting of the text-based data set introduced here in combination with radiology images of associated anatomic features - current work in the context of MEDICO on semantic annotation of anatomy images is described in (submitted).

Biomedical data sets[5] that are somewhat related to the one we are describing here include 'i2b2' on clinical data[6] next to the GENIA[7], BioText[8] and PASbio[9] corpora, all three of which have been designed for the extraction of terms/concepts and relations between them - see e.g. (Ciaramita et al., 2008) on relation extraction from the GENIA corpus. Importantly all of the corpora mentioned include manual annotation of term/concept/relation information, which is foreseen also in our work but is not yet discussed here.

Finally, more general work with Wikipedia as a data source is reported for instance in (Ruiz-Casado et al., 2006), (Strube and Ponzetto, 2006), (Völkel et al. 2006), all of which are more concerned with the extraction of

[4] http://www.nlm.nih.gov/research/visible/

[5] http://biocreative.sourceforge.net/bio_corpora_links.htm

[6] https://www.i2b2.org/NLP/

[7] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA

[8] http://biotext.berkeley.edu/data/dis_treat_data.html

[9] http://research.nii.ac.jp/~collier/projects/PASBio/

relation instances rather than relation types as is the topic or our work reported here.

3. Data Sources

In this section we describe the different data sources we used in setting up the human anatomy data set: the FMA ontology, the RadLex terminological resource and the Wikipedia pages on human anatomy. As the FMA was developed to fulfill the requirements of a formal ontology, whereas RadLex only a uniformed terminology, we will henceforth refer to the entities of the FMA ontology as *concepts* and to those of the RadLex as *terms*

3.1 Foundational Model of Anatomy (FMA)

As MEDICO targets semantic medical image retrieval, we use anatomical information to inform the system of concepts and spatial or partonomical relationships that are otherwise not obtained by the image parsing algorithms. The Foundational Model of Anatomy (FMA) that we use for this purpose is developed and maintained by the School of Medicine of the University of Washington and the US National Library of Medicine (Rosse and Mejino, 2003). Besides the specification of anatomy taxonomy, i.e. an inheritance hierarchy of anatomical entities, the FMA provides definitions for conceptual attributes, part-whole, location, and other spatial associations of anatomical entities. By additionally allowing for attributing relations (i.e. relations can be described in more detail by attaching additional attributes) FMA is particularly rich with respect to the specification of relations and, thus, can cope with the requirements for the precise and comprehensive capturing of the structure of the body.

FMA covers approximately 70,000 distinct anatomical concepts and more than 1.5 million relations instances from 170 relation types. The FMA is freely available as a Protégé 3.0 project or can be accessed via the Foundational Model Explorer [10]. There also exist conversions of the frame-based Protégé version of FMA to the OWL DL format (Goldbreich, Zhang, and Bodenreider, 2006), which version we use in the MEDICO project.

From the linguistic perspective the FMA ontology is complex in that the terms form cascaded structures. Often one term occurs within another such as in

Abdominal aorta
Abdominal aortic lumen
Abdominal aortic plexus
Abdominal aortic nerve plexus
…

Tunica media of abdominal aorta
Tunica intima of abdominal aorta
Lumen of abdominal aorta
…

A common structure is the following:

modifier [ANATOMICAL STRUCTURE]

where the *modifier* is one of the following:

modifier = {left, right, upper, lower, inferior, superior, lateral, anterior, anterolateral, antero-inferior, anteromedial, anterovential, posterior, ascending, descending, atrial, lower, upper}

as in

Left neck of mandible
Right neck of mandible
Anterior part of neck
Inferior part of back of neck
Lower lip skin
Upper trunk
Inferior ventricular vein
Superior vesical artery
Lateral aortic lymph node
Anterior body wall
Anterolateral central artery
Antero-inferior surface of body of pancreas
Anteromedial bronchial wall
Anteroventral nucleus
Posterior auricular vein
Ascending aorta
Descending aorta
Atrial cavity
Lower limb
Upper limb

Additionally, these modifiers share a common semantic characteristic in that they refer to anatomical locations. In this respect, the FMA is rich in morphological information that can be exploited to discover further domain relevant information such as the spatial information in anatomical locations.

Another example of linguistic richness (and source for ambiguity) is the use of prepositions such as *of.* In FMA terms, *of* is often used in the sense of *part_of* indicating a meronymy relationship as in *wall of pharynx.* However, this meronymy relationship cannot be assumed for each occurrence of *of,* as it can also occur in the sense of *matter of* or *tree of*. This complexity of FMA terms makes it hard to define regular patterns as exceptions occur quite often.

To address multilinguality, the FMA provides synonym information for almost every concept in languages as French, Italian, German, Spanish etc. However, also here ambiguities are present with respect to translations. For example, the German term *Öffnung* (*opening* in English) is listed as a synonym for the English term *mouth*. However, the term *Mund*, which is the actual German equivalent of *mouth* is not included.

Nevertheless, FMA is a valuable, and the most comprehensive machine-readable anatomical resource for medical information management and retrieval in the anatomy domain. Moreover, its complex terminological structure makes FMA a rich information source for linguistic analysis on our way to discover anatomy relevant relations and eventually to query patterns.

3.2 Radiology Lexicon (RadLex)

A major objective of the MEDICO project is to explore techniques for enhancing cross-modal medical image retrieval through automatic semantic annotation of

[10] http://fme.biostr.washington.edu:8089/FME/

domain-specific terms and relations. For this purpose, the publicly available Radiology Lexicon, or RadLex, is used. RadLex is a controlled vocabulary developed and maintained by the Radiological Society of North America (RSNA) for the purpose of uniform indexing and retrieval of radiology information, including images.

RadLex contains over 8,000 anatomic and pathologic terms, also those about imaging techniques, difficulties and diagnostic image qualities. It is available in a frame-based Protégé version.

As RadLex is thought of as a unified lexicon for capturing radiology information, it contains not only domain knowledge but also lexical information such as synonymy. Hence, entries such as *RadLex term, RadLex synonym* and *RadLex attribute* are present next to domain-specific information such as *drug-induced disorder*. Two different terms can be related to each other through various relationships. For example, at the lexical level the *Synonym of* relationship links the *Schatzki ring* and the *lower esophageal mucosal ring* to each other. Additionally, synonymy in natural language expressions is considered, so that the two expressions, *may be present* and *possibly present,* are also related to each other with the *Synonym of* relationship. Thus, using either one of these terms, it is possible to make assertions about the (un)certainty of an imaging observation. Examples of radiology specific relationships are *thickness of projected image* or *radiation dose.*

For the purpose of establishing a unified vocabulary, references are made to other medical controlled vocabularies such as the SNOMED-CT [11] via the *SNOMED_ID* or *SNOMED_Term* attributes. This way, RadLex enables the incorporation of the widely used SNOMED vocabulary in our relation extraction activities as an additional information resource Furthermore, (Marwede *et al*, 2007) reports on work about relating the RadLex terminology to the FMA by means of creating an intermediary ontology.

3.3 Wikipedia Anatomy Corpus

A central aspect of the query pattern mining task is the statistical analysis of the FMA and RadLex terms in relevant text collections, in order to assign relevance to a more precise set of terms and to investigate the most likely expressed (and hence queried) relations between them. For this purpose we need access to a representative corpus of anatomy texts, which could not be readily acquired as such data sets do not exist to our knowledge.

Instead, we selected the Wikipedia Category:Anatomy[12] as a starting point with 50 direct web pages and 20 sub-categories with 984 web pages in total. Out of these we selected about 900 web pages that are relevant to human anatomy. We removed for instance all web pages that are concerned with animal anatomy. Given the URLs for all selected web pages we were then able to generate and download an XML version of these using standard tools provided by Wikipedia[13]. As Wikipedia web pages consist of both text and structured data we then further analysed them to extract only the purely running text sections. In fact, we did not actually extract these sections but defined an extended XML format by which we were able to annotate which parts of the Wikipedia web page consists of running text vs. structured data. This step in the analysis left us with a corpus of around 500.000 tokens.

Finally, we ran all text sections through the TnT part-of-speech parser (Brants, 2000) to extract all nouns in the corpus and to compute a relevance score (chi-square) for each by comparing anatomy frequencies with those in the British National Corpus (BNC)[14]. This information in turn allowed us to compute the relevance of each sentence in the corpus, which we intend to use as a further focus in the relation extraction task. A next step will be to parse all sentences and annotate them with predicate-structure information, which may be then used for relation extraction along the lines of (Schutz and Buitelaar, 2005).

4. Data processing

In this section we describe the data processing steps in using the FMA ontology and RadLex terminological resource to identify the most relevant anatomy and radiology concepts in the Wikipedia anatomy corpus.

These two external knowledge resources provide valuable and publicly available domain information in a very specific and sensitive domain as medicine and especially when the clinical expert knowledge is not available or not easily accessible. Moreover, as these two resources are developed together by domain experts and computer scientists, they are reliable, hence suit our purposes well.

Furthermore, extracted terms from each resource are compared and aligned to obtain an integrated radiology view of human anatomy, on which we intend to focus the relation extraction and query pattern mining task. Data processing for each resource both for binary mappings between the two resources and for those with the Wikipedia was done based on exact term matching. Currently, no term variations or synonyms have been considered but remains as future work.

4.1 FMA Concepts and Relations

The concepts and relationships from the FMA ontology are used to identify the human anatomy relevant terms and relationships from the Wikipedia anatomy corpus.

As a first step the concepts and the relationships from the FMA ontology were extracted yielding a list of 124,769 entries. However, this list includes very generic terms such as *Anatomical structure* as well as very specific terms such as *Anastomotic branch of right anterior inferior cerebellar artery with right superior cerebellar artery*. Generic terms were filtered out from the list as they will not deliver sufficient information because any human body part can be an anatomical structure. Similarly, very specific terms are also filtered out as they most likely will not occur in the Wikipedia corpus (or in any other corpus for that matter). After filtering such terms, the resulting list consists of 19,367 terms (consisting of one, two or three words) such as:

[11] http://www.snomed.org/

[12] http://en.wikipedia.org/wiki/Category:Anatomy

[13] http://en.wikipedia.org/wiki/Special:Export

[14] The BNC (http://www.natcorp.ox.ac.uk/) is a 100 million word collection of samples of written and spoken language from a wide range of sources, designed to represent a wide cross-section of current British English.

abdominal lymph node
femoral head
jugular lymphatic trunk
left radial neck
ligament of neck
lymph node
lymph node capsule
…

The statistically most relevant FMA terms were identified on the basis of chi-square scores computed for nouns in the anatomy corpus (see section 3.3 above). Single word terms in the FMA and occurring in the corpus correspond directly to the noun that the term is build up of (e.g. the noun 'ear' corresponds to the FMA term *ear*). In this case, the statistical relevance of the term is the chi-square score of the corresponding noun. In the case of multi-word terms occurring in the corpus, the statistical relevance is computed on basis of the chi-square score for each constituting noun and/or adjective in the term, summed and normalized over the length of the term, e.g. the relevance value for *lymph node* is the summation of the chi-square scores for 'lymph' and 'node' divided by 2.

In order to take frequency in account, we further multiplied the summed relevance value by the frequency of the term. This assures that only frequently occurring terms are judged as relevant. A selection (top 20) of the resulting list of most relevant FMA terms is shown in Table 1 below:

FMA Terms	**Relevance**
Lateral	109407852
Anterior	76204800
Muscle	69101264
Posterior	36516690
Medial	33953121
Artery	27914139
Cortex	21314304
Dorsal	19520100
Inferior	16855128
Superior	14028800
Deep	9763904
Central	8763650
Internal	7883937
Duct	7659345
Bone	6976292
Membrane	6202856
Embryo	5469423
Ligament	5372510
Organ	5119666
Gland	4480047

Table 1: Top 20 of statistically most relevant FMA terms in the anatomy corpus

We then further studied the context of selected relevant FMA terms, i.e. the sentences in which they occurred in the anatomy corpus. For instance, the term *cervical lymph node* occurs in the Wikipedia page on cervical lymph nodes[15] as follows: …*cervical lymph nodes are lymph nodes found in the neck*. This information is of importance as it delivers spatial information on the location of cervical lymph nodes (i.e. in the neck), which is not trivial for image parsing and pattern recognition algorithms to obtain automatically. Such an observation motivates a statistical and linguistic analysis of the terms *cervical lymph nodes*, *neck* and the verb *found* to determine if their coocurrence could be formalized as the *located_in* relation between these two concepts and whether this is a common pattern. If so, a corresponding query pattern hypothesis can be generated as follows:

cervical lymph nodes	*located_in*	[HEAD NECK REGION]
	AND	
[HEAD NECK REGION]	*has_part*	*neck*

After evaluation by the clinical experts, the statistical and linguistic analysis can be tuned to obtain the required level of granularity.

4.2 RadLex Concepts and Relations

Following a similar approach, RadLex was used to identify the most relevant radiology terms in the anatomy corpus. RadLex includes only the radiology relevant part of human anatomy knowledge, therefore the RadLex terminology is more lightweight compared to that of the FMA. Complex morphological information and cascaded structures are seldom.

An initial list of terms that consisted of 13,156 entries was extracted from the RadLex controlled vocabulary by parsing the downloadable version from the project web site. After manual clean up, such as removing the duplicates, the list was reduced to a subset of 12,055 entries. In contrast to the FMA, the resulting term list contains also very specific terms, longer than 3 words. It was not necessary to exclude them as there are only few and keeping them did not particularly increase the size of the term list to cause efficiency problems.

One-word terms however require an additional step of handling abbreviations, which is currently not addressed in our work. Since Radlex is designed as a unifying resource it contains a wide range of domain specific abbreviations from pathology, physiology, radiology etc. Abbreviation resolution in the biomedical domain is an active research field on its own – see e.g. (Chang et al., 2002), (Yu et al., 2004), (Schwartz and Hearst, 2003).

The RadLex vocabulary includes several types of relationships, e.g. *is a*, *part of*, *contained in*, *continuous with*. These relationships are included as terms in the list as well as radiology specific relationships such as *"advanced analytic difficulty"*. Table 2 shows results for the statistical relevance analysis for RadLex terms:

[15] http://en.wikipedia.org/wiki/Cervical_lymph_nodes

RadLex Terms	Relevance
Anterior	228614400
Lateral	109407852
Posterior	73033380
First	47752592
Small	45871353
Large	34303500
Medial	33953121
Tissue	32979809
Artery	27914139
Cortex	21314304
Dorsal	19520100
Inferior	16855128
Superior	14028800
Body	10360776
Deep	9763904
Long	9604980
Brain	8765497
Central	8763650
Major	8696892
Blood	8655100

Table 2: Top 20 of statistically most relevant RadLex terms in the anatomy corpus

Studying the corpus context of the RadLex term *lymphoma* showed that it frequently occurs with terms such as *Hodgkin's Disease* and *Non-Hodgkin lymphoma* as in: … *lymphoma (previously called Hodgkin's Disease) and Non-Hodgkin lymphoma* [16] . These observations, again, motivate a further linguistic and statistical analysis to see if a structural query pattern can be discovered for these terms and relations between them. For example we may hypothesize that *Hodgkin lymphoma* and *Non-Hodgkin lymphoma* are common lymphoma types that clinicians would typically search for.

4.3 Correspondence between RadLex and FMA

In the process of preparing the human anatomy data set with a radiology perspective, it was necessary to identify the correspondences between the two vocabularies. As both vocabularies are designed for different purposes, their range and granularity are also different. Consequently, the FMA has a much more detailed view of human anatomy, which may result in information overload for the purpose of image annotation. The RadLex vocabulary on the other hand, does cover anatomical information but may not suffice occasionally. Nevertheless, it includes image specific information, such as image quality, which is not typically found in an anatomy ontology. The image quality information can be useful for image annotation scenarios, where the radiologist wants to filter out all images whose resolution fall below a certain threshold. In order to bring out the best of both, we started with a comparison of both vocabularies at the lexical level.

Accordingly, the two (subset) term lists, namely those of the RadLex and the FMA were compared to each other to identify the intersections. As a result, 1259 perfect matches were identified, i.e. they were identical in both terminologies. Some examples of these terms are *abdominal lymph node, adrenal cortex, foramen magnum* etc.

The comparison results showed that often one RadLex term occurs within an FMA term as in:

RadLex term:	*iliac crest*
FMA term:	*left iliac crest*
FMA term:	*left iliac crest tubercle*

The reason for this is twofold. Firstly, as discussed, the level of granularity in the FMA is much finer than that of RadLex. Hence, there can be multiple terms to describe one anatomical structure. Secondly, as mentioned earlier, linguistically the FMA terms have a much more complex structure. This can be observed also in the following case:

RadLex term:	*arcuate nucleus*
FMA Term:	*arcuate nucleus-1*
FMA term:	*arcuate nucleus-2*
FMA term:	*arcuate nucleus-3*

This situation has also consequences on the semantic level in that the mapping between RadLex and corresponding FMA concepts becomes ambiguous. As a first step to address these problems, the terms of both vocabularies are ranked according to simple string similarity – here we build on work described in (submitted). More sophisticated, semantic similarity measures do exist for the biological domain (Pedersen et al, 2007) that can as next be incorporated for the purposes of this work.

An example of such a mapping between RadLex and FMA terms is shown in Table 3 below. The numbers in the third column show the simple string distance according to different characters.

RadLex Term	FMA Term	String Distance
anconeus	subanconeus	3
	left anconeus	5
	right anconeus	6
	nerve to anconeus	9
	fascia of anconeus	10

Table 3: RadLex term with list of corresponding FMA terms according to string distance

5. Conclusions and Future Work

We described our ongoing work towards establishing a data set for human anatomy, which we intend to use for relation extraction and query pattern mining in the context of semantic retrieval of radiology images.

In a next phase we will extend the current data set with related scientific abstracts that can be obtained from PubMed. Here we intend to exploit our analysis of term relevance as discussed in this paper, by selecting primarily PubMed articles related to the top most relevant terms (FMA/RadLex) that we identified in the Wikipedia anatomy corpus.

[16] http://en.wikipedia.org/wiki/Ann_Arbor_staging

We realize that the most relevant data resource for the detection of potentially relevant query patterns may in fact be clinical protocols (patient records) rather than anatomy textbook articles as found on Wikipedia or even scientific abstracts from PubMed. Typically, patient records include information about image findings, diagnosis, symptoms, disease codes etc. On the downside however, they are very difficult to obtain because of data protection and confidentiality matters. Consequently, we therefore decided to consult open source knowledge resources such as Wikipedia and PubMed in the initial phase, but we will be able to obtain a relevant set of patient records during the course of the project.

After all of the knowledge sources are in place however, we will be able to identify a coherent picture of anatomy concepts in the context of radiology images and their relations as expressed in biomedical theory (Wikipedia, PubMed) and clinical practice (PubMed, patient records).

Acknowledgments

This research has been supported in part by the THESEUS Program in the MEDICO Project, which is funded by the German Federal Ministry of Economics and Technology under the grant number 01MQ07016. The responsibility for this publication lies with the authors. We thank Thomas Eigner from the DFKI Language Technology Lab for implementing parts of the system.

References

Brants T. (2000). TnT - A Statistical Part-of-Speech Tagger. In: *Proc. of the 6th ANLP Conference*, Seattle, WA

Chang, J.T., Schütze, H., Altman, R.B. (2002). Creating an online dictionary of abbreviations from medline. *J Am Med Inform Assoc* 9(6), pp.612-620.

Ciaramita, M., Gangemi, A., Ratsch, E., Saric, J., Rojas, I. (2008). Unsupervised Learning of Semantic Relations for Molecular Biology Ontologies. In Paul Buitelaar, Philipp Cimiano (Eds.) *Ontology Learning and Population: Bridging the Gap between Text and Knowledge*. Frontiers in Artificial Intelligence and Applications Series, Vol. 167, IOS Press.

Langlotz, CP. (2006). RadLex: A New Method for Indexing Online Educational Materials In: *Radiographics* 26, pp.1595-1597.

Pedersen T, Pakhomov V, Patwardhan S and Chute C. G. (2007) Measures of semantic similarity and relatedness in the biomedical domain. *Journal of Biomedical Informatics*, 40(3) pp. 288-299.

Marwede D., Fielding M. and Kahn T. (2007). *RadiO: A Prototype Application Ontology for Radiology Reporting Tasks*. In *Proceedings of the AMIA 2007 Symposium*, Chiocago IL.

Rosse C. and J.L.V. Mejino Jr. (2003). A reference ontology for biomedical informatics: the foundational model of anatomy. *Journal of Biomedical Informatics*, 36(6), pp. 478–500.

Ruiz-Casado M., E. Alfonseca, and P. Castells. (2006). From Wikipedia to Semantic Relationships: a Semi-automated Annotation Approach. In *Workshop on Semantic Wikis at the 3rd European Semantic Web Conference*, Montenegro.

Schreiber G., Akkermans H., Anjewierden A., Dehoog R., Shadbolt N., Vandevelde W., and Wielinga B. (1999). *Knowledge Engineering and Management: The CommonKADS Methodology*. MIT Press.

Schutz A., and P. Buitelaar. (2005). RelExt: A Tool for Relation Extraction in Ontology Extension In: *Proceedings of the 4th International Semantic Web Conference*, Galway, Ireland.

Schwartz, A, Hearst, M, (2003). A simple algorithm for identifying abbreviation definitions in biomedical texts. In: *Proceedings of the Pacific Symposium on Biocomputing*, Kauai, Hawaii.

Strube M., and S. Ponzetto. (2006). WikiRelate! Computing Semantic Relatedness Using Wikipedia. In: *Proceedings of the 21st National Conference on Artificial Intelligence*, Boston, MA.

Sure Y., Staab. S., and Studer R. (2002). Methodology for Development and Employment of Ontology based Knowledge Management Application. In *SIGMOD Special Issue on Semantic Web, Database Management and Information Systems*, 31(4) pp.18-23.

Yu H, Kim, W., Hatzivassiloglou, V and Wilbur, W.J.J. (2004). Using MEDLINE as a knowledge source for disambiguating abbreviations and acronyms in full-text biomedical journal articles. In *Proceedings 17th IEEE Symposium On Computer-Based Medical Systems*, Bethesda, MD.