

Automatically Extracting Enzyme Interaction and Protein Structure Information from Biological Science Journal Articles

Kevin Humphreys ; George Demetriou ; Robert Gaizauskas

Department of Computer Science, University of Sheffield

Regent Court, Portobello Street

Sheffield S1 4DP UK

{kwh ; demetri ; robertg}@dcs.shef.ac.uk

Abstract

With the explosive growth of scientific literature in the area of molecular biology, the need to automatically process and extract information from on-line text sources has become increasingly important. In this paper we consider the application of Information Extraction (IE) technology to the extraction of factual information from biological journal papers. IE has proved successful at extracting information primarily from newswire texts and primarily in domains concerned with human activity as demonstrated by the systems that took part in the U.S. DARPA Message Understanding Conferences (MUCs). We describe how an information extraction system designed to participate in the MUC exercises has been modified for two bioinformatics applications: EMPATHIE, concerned with enzyme and metabolic pathways; and PASTA, concerned with protein structure. The progress so far provides convincing grounds for believing that IE techniques will deliver novel and effective ways for the extraction of information from unstructured text sources in the pursuit of knowledge in the biological domain.

1 Introduction

Information Extraction (IE) may be defined as the activity of extracting details of predefined classes of entities and relationships from natural language texts and placing this information into a structured representation called a *template* (Cowie and Lehnert, 1996; Gaizauskas and Wilks, 1998). The prototypical IE tasks are those defined by the U.S. DARPA-sponsored Message Understanding Conferences (MUCs), requiring the filling of a complex template from newswire texts on subjects such as joint venture announcements, management succession events, or rocket launchings (Def, 1995, 1998). While the performance of current technology is not yet at human levels overall, it is approaching human levels for some component tasks (e.g. the recognition and classification of named entities in text) and is at a level at which comparable technologies, such as information retrieval and machine translation, have found useful application. IE is particularly relevant where large volumes of text make human analysis infeasible, where template-oriented information seeking is appropriate (i.e. where there is a relatively stable information need and a set of texts in a relatively narrow domain), where conventional information retrieval technology is inadequate, and where some error can be tolerated.

One area where we believe these criteria are met, and where IE techniques have as yet been applied only in a limited way (though see Fukuda et al. (1998); Rindfleisch et al. (2000); Thomas et al. (2000)), is the construction

of databases of scientific information from journal articles, for use by researchers in molecular biology. The explosive growth of textual material in this area means that no one can keep up with what is being published. Conventional retrieval technology returns both too little, because of the complex, non-standardised terminology in the area, and too much, because what is sought is not whole texts in which key terms appear, but facts buried in the texts. Further, useful templates can be defined for some scientific tasks. For example, scientists working on drug discovery have an ongoing interest in reactions catalysed by enzymes in metabolic pathways. These reactions may be viewed as a class of events, like corporate management succession events, in which various classes of entities (enzymes, compounds) with attributes (names, concentrations) are related by participating in the event in specific roles (substrate, catalyst, product). Finally, some error can be tolerated in these applications, because scientists can verify the information against the source texts – the technology serves to assist, not to replace, investigation.

Thus, we believe automatically extracting information from scientific journal papers is an important and feasible application of IE techniques. It is also interesting from the perspective of IE research because it extends IE to *domains* and to *text genres* where it has never been applied before. To date most IE applications have been to domains of human activity, predominately economic activity, and have involved newswire texts which have a characteristic lexis, structure and length. Applying IE to

scientific journal papers in the area of molecular biology means a radical shift of subject domain away from the world of people, companies, products and places that have largely figured in previous applications. It also means dealing with a text genre in which there is a vast and complex technical vocabulary, where the texts are structured into subsections dealing with method, results, and discussion, and where the texts are much longer. These differences all pose tough challenges for IE techniques as developed so far: can they be applied successfully in this area?

In this paper we describe the use of the technology developed through MUC evaluations in two bioinformatics applications. The next section describes the general functionality of an IE system, and section 3 then describes the two specific applications on which we are working: extraction of information about enzymes and metabolic pathways and extraction of information about protein structure, in both cases from scientific abstracts and journal papers. Section 4 describes the principle processing stages and techniques of our system, and section 5 presents evaluations of the system's performance. While much further refinement of the system for both applications is possible, indications are that IE can indeed be successfully applied to the task of extracting information from scientific journal papers.

2 Information Extraction Technology

The most recent MUC evaluation (MUC-7, (Def, 1998)) specified five separate component tasks, which illustrate the main functional capabilities of current IE systems:

1. *Named Entity recognition* requires the recognition and classification of named entities such as organisations, persons, locations, dates and monetary amounts.
2. *Coreference resolution* requires the identification of expressions in the text that refer to the same object, set or activity. These include variant forms of name expression (*Ford Motor Company ... Ford*), definite noun phrases and their antecedents (*Ford ... the American car manufacturer*), and pronouns and their antecedents (*President Clinton ... he*). Coreference relations are only marked between certain syntactic classes of expressions (noun phrases and pronouns) and a relatively constrained class of relationships to mark is specified, with clarifications provided with respect to bound anaphors, apposition, predicate nominals, types and tokens, functions and function values, and metonymy.
3. *Template Element filling* requires the filling of small scale templates (slot-filler structures) for specified classes of entity in the texts, such as organisations, persons, certain artifacts, and locations, with slots

such as name (plus name variants), description as supplied in the text, and subtype.

4. *Template Relation filling* requires filling a two slot template representing a binary relation with pointers to template elements standing in the relation. For example, a template relation of `employee_of` containing slots for a person and organisation is filled whenever a text makes clear that a particular person is employed by a particular organisation. Other relations are `product_of` and `location_of`.
5. *Scenario Template filling* requires the detection of relations between template elements as participants in a particular type of event, or scenario (rocket launches for MUC-7), and the construction of an object-oriented structure recording the entities and various details of the relation.

Systems are evaluated on each of these tasks as follows. Each task is precisely specified by means of a task definition document. Human annotators are then given these definitions and use them to produce by hand the 'correct' results for each of the tasks – filled templates or texts tagged with name classes or coreference relations (these results are called *answer keys*). The participating systems are then run and their results, called *system responses*, are automatically scored against the answer keys. Chief metrics are *precision* – percentage of the system's output which is correct (i.e. occurs in the answer key) – and *recall* – percentage of the correct answer which occurs in the system's output.

State-of-the-art (MUC-7) results for these five tasks are as follows (in the form recall/precision): named entity – 92/95; coreference – 56/69; template element – 86/87; template relation – 67/86; scenario template 42/65.

3 Two Bioinformatics Applications of Information Extraction

We are currently investigating the use of IE for two separate bioinformatics research projects. The Enzyme and Metabolic Pathways Information Extraction (EMPathIE) project aims to extract details of enzyme reactions from articles in the journals *Biochimica et Biophysica Acta* and *FEMS Microbiology Letters*. The utility for biological researchers of a database of enzyme reactions lies in the ability to search for potential sequences of reactions, where the products of one reaction match the requirements of another. Such sequences form metabolic pathways, the identification of which can suggest potential sites for the application of drugs to affect a particular end result. Typically, journal articles in this domain describe details of a single enzyme reaction, often with little indication of related reactions and which pathways the reaction may be part of. Only by combining details from several articles can potential pathways be identified.

The Protein Active Site Template Acquisition (PASTA) project aims to extract information concerning the roles of amino acids in protein molecules, and to create a database of protein active sites from both scientific journal abstracts and full articles. The motivation for the PASTA project stems from the need to extract and rationalise information in the protein structure literature. New protein structures are being reported at very high rates and the number of co-ordinate sets (currently about 12000) in the Protein Data Bank (PDB) (Bernstein et al., 1977) can be expected to increase ten-fold in the next five years. The full evaluation of the results of protein structure comparisons often requires the investigation of extensive literature references, to determine, for instance, whether an amino acid has been reported as present in a particular region of a protein, whether it is highly conserved, implicated in catalysis, and so on. When working with several different structures, it is frequently necessary to go through a large number of scientific articles in order to discover any functional or structural equivalences between residues or groups of residues. Computational methods that can extract information directly from these articles would be very useful to biologists in comparison classification work and to those engaged in modelling studies.

The following section describes the EMPathIE and PASTA tasks, including the intended extraction results from documents containing text such as that shown in Figure 1.

Results: We have determined the crystal structure of a triacylglycerol lipase from Pseudomonas cepacia (Pet) in the absence of a bound inhibitor using X-ray crystallography. The structure shows the lipase to contain an alpha/beta-hydrolase fold and a catalytic triad comprising of residues Ser87, His286 and Asp264. The enzyme shares several structural features with homologous lipases from Pseudomonas glumae (PgL) and Chromobacterium viscosum (CvL), including a calcium-binding site. The present structure of Pet reveals a highly open conformation with a solvent-accessible active site, This is in contrast to the structures of PgL and Pet in which the active site is buried under a closed or partially opened 'lid', respectively.

Figure 1: Sample Text Fragment from a Scientific Paper in Molecular Biology

3.1 EMPathIE

One of the inspirations for the Enzyme and Metabolic Pathways application was the existence of a manually constructed database for the same application. The EMP database (Selkov et al., 1996) contains over 20,000 records of enzyme reactions, collected from journal articles published since 1964. That such a database has been constructed and is widely used demonstrates the utility of the application. EMPathIE aims to extract only a key subset of the fields found in the EMP database records.

The main fields required in a record of an enzyme reaction are: the enzyme name, with an enzyme classification (EC) number, if available, the organism from which the enzyme was extracted, any known pathway in which the reaction occurs, compounds involved in the reaction, with their roles classified as either substrate (input), product (output), activator, inhibitor, cofactor or buffer, and any compounds known not to be involved in the reaction, with their roles classified as either non-substrate or non-product.

The template definitions include three Template Elements: *enzyme*, *organism* and *compound*, a single Template Relation: *source*, relating *enzyme* and *organism* elements, and a Scenario Template for the specific metabolic pathway task. The Scenario Template describes a pathway involving one or more interactions, each of which is a reaction between an enzyme and one or more participants, possibly under certain constraints. A manually produced sample Scenario Template is shown here, taken from an article on 'isocitrate lyase activity' in FEMS Microbiology Letters.

```
<ENZYME-1> :=
  NAME: isocitrate lyase
  EC_CODE: 4.1.3.1

<ORGANISM-1> :=
  NAME: Haloferax volcanii
  STRAIN: ATCC 29605
  GENUS: halophilic Archaea

<COMPOUND-1> :=
  NAME: phenylhydrazone

<COMPOUND-2> :=
  NAME: KCl

<SOURCE-1> :=
  ENZYME: <ENZYME-1>
  ORGANISM: <ORGANISM-1>

<PATHWAY-1> :=
  NAME: glyoxylate cycle
  INTERACTION: <INTERACTION-1>

<INTERACTION-1> :=
  ENZYME: <ENZYME-1>
  PARTICIPANTS: <PARTICIPANT-1>
                <PARTICIPANT-2>

<PARTICIPANT-1> :=
  COMPOUND: <COMPOUND-1>
  TYPE: Product
  TEMPERATURE: 35C

<PARTICIPANT-2> :=
  COMPOUND: <COMPOUND-2>
  TYPE: Activator
  CONCENTRATION: 1.75 M
```

This template describes a single interaction found to

be part of the metabolic pathway known as the *glyoxylate cycle*, where the interaction is between the enzyme *isocitrate lyase* and two other participants. The first participant is the compound *glyoxylate phenylhydrazone*, which has the role of a *product* of the interaction at a temperature of 35C. The second is the compound *KCl*, which has the role of an *activator* at a concentration of 1.75M.

The template design follows closely the MUC-style IE template, and is richer than the EMP database record format in terms of making relationships between entities explicit. However, most of the slot values can still be mapped back to the EMP format to allow an automatic evaluation of system output against the manually constructed EMP resource.

3.2 PASTA

The entities to be extracted for the PASTA task include proteins, amino acid residues, species, types of structural characteristics (secondary structure, quaternary structure), active sites, other (probably less important) regions, chains and interactions (hydrogen bonds, disulphide bonds etc.) In collaboration with molecular biologists we have designed a template to capture protein structure information, a fragment of which, filled with information extracted from the text in Figure 1, is shown below:

```
<RESIDUE-str97-521>:=
  RESIDUE_TYPE:  SERINE
  RESIDUE_NO:    "87"
  IN_PROTEIN:   <PROTEIN-str97-521>
  SITE/FUNCTION: "active site"
                "catalytic"
                "interfacial activation"
                "calcium-binding site"
  SECOND_STRUCT: alpha-helix
  REGION:       'lid'
  ARTICLE:     <ARTICLE-str97-521>

<PROTEIN-str97-521>:=
  NAME:         "Triacylglycerol lipase"
  SCOP_CLASS:  "Lipase"
  PDB_CODE:    1LGY
  IN_SPECIES:  <SPECIES-str97-521>

<SPECIES-str97-521>:=
  NAME:        "Pseudomonas cepacia"
  NAME_TYPE:   SCIENTIFIC
```

The residue information contains slots that describe the structural characteristics of the particular protein (e.g. SECONDARY structure, REGION) and the importance of the residue in the structure (e.g. SITE/FUNCTION). Other slots serve as pointers, linking different template objects together to represent relational information between entities (e.g. the IN_PROTEIN and IN_SPECIES slots). Further Template Relations can also be defined to link proteins or residues with structural equivalence.

4 The EMPATHIE and PASTA Systems

The IE systems developed to carry out the EMPATHIE and PASTA tasks are both derived from the Large Scale Information Extraction (LaSIE) system, a general purpose IE system, under development at Sheffield since 1994 (Gazauskas et al., 1995; Humphreys et al., 1998). One of several dozen systems designed to take part in the MUC evaluations over the years, the LaSIE system more or less fits the description of a generic IE system (Hobbs, 1993). LaSIE is neither as ‘deep’ as some earlier IE systems that attempted full syntactic, semantic and discourse processing (Hobbs, 1991) nor as ‘shallow’ as some recent systems that use finite state pattern matching techniques to map directly from source texts to target templates (Appelt et al., 1995). The processing modules which make up the EMPATHIE system are shown in figure 2, within the GATE development environment (Cunningham et al., 1997). The PASTA system is similar and reuses several modules, within the same environment. The architecture of the original LaSIE system has been substantially rearranged for its use in the biochemical domain, mainly to allow the reuse of general English processing modules, such as the part-of-speech tagger and the phrasal parser, without special retraining or adaptation to allow for the domain-specific terminology. This has resulted in an independent terminology identification subsystem, postponing general syntactic analysis until an attempt to identify terms has been made. In general, the original LaSIE system modules, developed for newswire applications, have been reused, but with various modifications resulting from specific features of the texts, as described in the following. Both systems have a pipeline architecture consisting of four principal stages, described in the following sections: *text preprocessing* (SGML/structure analysis, tokenisation), *lexical and terminological processing* (terminology lexicons, morphological analysis, terminology grammars), *parsing and semantic interpretation* (sentence boundary detection, part-of-speech tagging, phrasal grammars, semantic interpretation), and *discourse interpretation* (coreference resolution, domain modelling).

4.1 Text Preprocessing

Scientific articles typically have a rigid structure, including abstract, introduction, method and materials, results, and discussion sections, and for particular applications certain sections can be targeted for detailed analysis while others can be skipped completely. Where articles are available in SGML with a DTD, an initial module is used to identify particular markup, specified in a configuration file, for use by subsequent modules. Where articles are in plain text, an initial ‘sectioniser’ module is used to identify and classify significant sections using sets of regular expressions. Both the SGML and sectioniser modules may specify that certain text regions are to be ex-

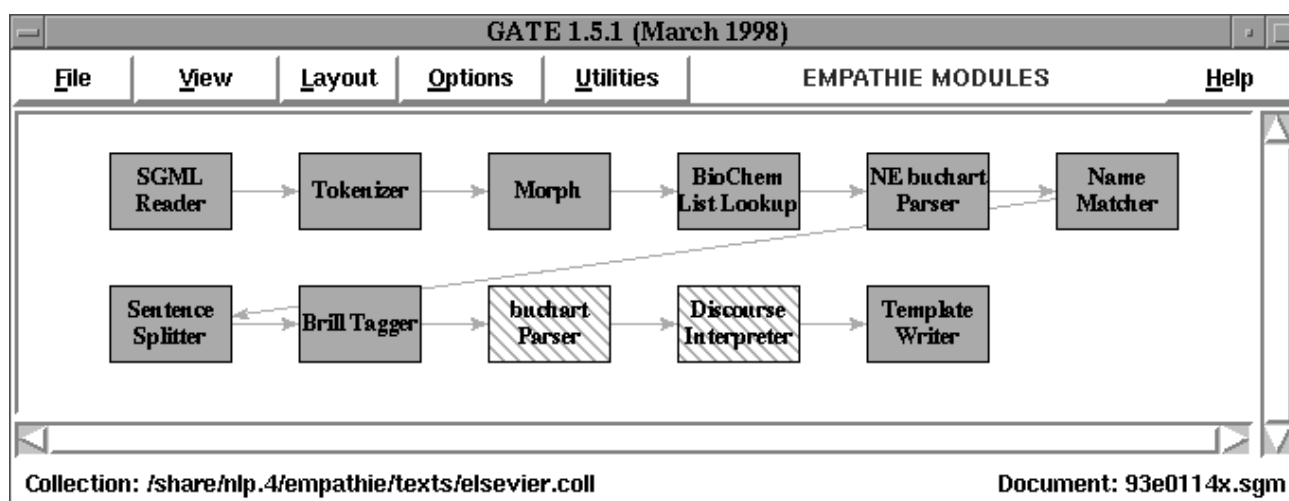


Figure 2: EMPATHIE system modules within GATE

cluded from any subsequent processing, avoiding detailed processing of apparently irrelevant text, especially within the discourse interpretation stage where coreference resolution is a relatively expensive operation.

The tokenisation of the input needs to identify tokens within compound names, such as abbreviations like *NaCl*, where *Na* and *Cl* need to be matched separately in the lexical lookup stage to avoid listing all possible sequences explicitly. The tokenisation module must therefore make as few assumptions as possible about the input, proposing minimal tokens which may be recombined in subsequent stages.

4.2 Lexical and Terminological Processing

The main information sources used for terminology identification in the biochemical domain are: case-insensitive terminology lexicons, listing component terms of various categories; morphological cues, mainly standard biochemical suffixes; and hand-constructed grammar rules for each terminology class. For example, the enzyme name *mannitol-1-phosphate 5-dehydrogenase* would be recognised firstly by the classification of *mannitol* as a potential compound modifier, and *phosphate* as a compound, both by being matched in the terminology lexicon. Morphological analysis would then suggest *dehydrogenase* as a potential enzyme head, due to its suffix *-ase*, and then grammar rules would apply to combine the enzyme head with a known compound and modifier which can play the role of enzyme modifier.

The biochemical terminology lexicons, acquired from various publicly available resources, have been structured to distinguish various term components, rather than complete terms, which are then assembled by grammar rules. Resources such as the SWISS-PROT list of official enzyme names were manually split into separate lists of component terms, based purely on their apparent syntactic

structure rather than any expert knowledge of whatever semantic structure the names reflect. Corresponding grammar rules were then added to recombine the components. Of course, lists of complete multi-word terms can also be used directly in the lexicons, but the rule-based approach has the advantage of being able to recognise novel combinations, not explicitly present in the term lists, and avoids reliance on the accuracy and completeness of available terminology resources. Component terms may also play multiple roles in different terminology classes, for instance amino acid names may be components of both protein and enzyme names, as well as terms in their own right, but the rule-based approach to terminology recognition means they only need to be listed in a single terminology category. The total number of terminology lexicon entries for the biochemical terms is thus comparable to other domains, with approximately 25,000 component terms at present in 52 categories.

4.3 Parsing and Semantic Interpretation

The syntactic processing modules treat any terms recognised in the previous stage as non-decomposable units, with a syntactic role of proper noun. The sentence splitting module cannot therefore propose sentence boundaries within a preclassified term. Similarly, the part-of-speech tagger only attempts to assign tags to tokens which are not part of proposed terms, and the phrasal parser treats terms as preparsed noun phrases. Of course, this approach does not necessarily assume the terminology recognition subsystem to be fully complete and correct, and subsequent syntactic or semantic context can still be used to reclassify or remove proposed terms. In particular, tokens which are constituents of terms proposed but not classified by the NE subsystem, i.e. potential but unknown NEs, are passed to the tagger and phrasal parser as normal, but the potential term is passed to the parser in

addition, as a proper name, to allow the phrasal grammar to determine the best analysis. If the unclassified NE is retained after phrasal parsing, it may be classified within the discourse interpreter, using its semantic context or as a result of being coreferred with an entity of a known class.

The phrasal grammar includes compositional semantic rules, which are used to construct a semantic representation of the ‘best’, possibly partial, parse of each sentence. This predicate logic-like representation is passed on as input to the discourse interpretation stage.

4.4 Discourse Interpretation

The discourse interpreter adds the semantic representation of each sentence to a predefined *domain model*, made up of an ontology, or concept hierarchy, plus inheritable properties and inference rules associated with concepts. The domain model is gradually populated with instances of concepts from the text to become a *discourse model*. A powerful coreference mechanism attempts to merge each newly introduced instance with an existing one, subject to various syntactic and semantic constraints. Inference rules of particular instance types may then fire to hypothesise the existence of instances required to fill a template (e.g. an organism with a *source* relation to an enzyme), and the coreference mechanism will then attempt to resolve the hypothesised instances with actual instances from the text.

The template writer module reads off the required information from the final discourse model and formats it as in the template specification.

Initial domain models for the EMPATHIE and PASTA tasks have been manually constructed directly from the template definition. This involves the addition of concept nodes to the system’s semantic network for each of the entities required in the template, with subhierarchies for possible subtypes, as required. Property types are added for each of the template slots (e.g. *concentration*, *temperature*), and consequence rules added to hypothesise instances for each slot of a template entity, from an appropriate textual trigger. The Discourse Interpreter’s general coreference mechanism is then used to attempt to resolve hypothesised instances with instances mentioned in the text. Subsequent refinement of these models will involve extending the concept subhierarchies and the addition of coreference constraints on the hypothesised instances, based on available training data.

5 Results and Evaluation

5.1 Evaluation

Currently, a complete prototype EMPATHIE system exists which can produce filled templates as specified above. This prototype has been developed by concentrating on the full texts of six journal papers (the *development* corpus) and evaluated against a corpus of a further seven

journal papers (the *evaluation* corpus). Filled templates for all thirteen of these journal papers were produced by trained biochemists highlighting key entities on paper copies of the texts and adding marginal notes where necessary to specify compound roles in interactions and any additional slot values such as concentration, temperature, etc. The annotations were translated to template format by the system developer (with the system frozen before evaluation texts were seen), but some degree of subjective interpretation was required in this process. The annotation would therefore probably be difficult to reproduce without a detailed task specification document, which would be aided by inter-annotator agreement studies to highlight areas of ambiguity in the task definition. However, the current templates at least have the advantage of being produced with some degree of consistency by the developer alone, and so do allow a useful measure of the system’s accuracy.

Overall template filling results are shown in Table 1. The columns show: the number of items the system correctly identified (CORrect), the number of items where the system response and the answer key differed (INCORrect), the number of items the system missed (MISSing), the number the system spuriously proposed (SPURious) and the standard metrics of RECall and PREcision, discussed in section 2 above. Here “items” refers to filled slot occurrences in the templates. Scoring proceeds by first aligning template objects in the system response with objects in the answer key and then counting the number of matching slot fills in the aligned objects (see Def (1998) for details).

Test Set	COR	INC	MIS	SPU	REC	PRE
Dev	150	121	330	61	25	45
Eval	213	193	518	93	23	43

Table 1: Initial Template results for EMPATHIE

In addition to evaluating the template filling capabilities of the prototype we have evaluated its performance at correctly identifying and classifying term classes in the texts (this corresponds to the MUC named entity task). To do this six of the seven evaluation corpus articles were manually annotated for eleven terminology or named entity classes. The results are shown in Table 2¹.

The PASTA system has been implemented as far as the terminology recognition stage. Preliminary template design, as indicated above, has been carried out, and we are starting to build a domain model. A corpus of 52 abstracts of journal articles has been manually annotated with terminology classes, by the system developer with

¹In calculating both EMPATHIE and PASTA terminology results we have used a weak criterion of correctness whereby a response is correct if its type matches the type of the answer key and its text extent matches a substring of the key’s extent. Insisting on the stronger matching criterion of strict string identity lowers recall and precision scores by approximately 4 % overall

Name_Type	COR	INC	MIS	SPU	REC	PRE
compound	100	6	156	3	38	92
element	22	0	17	0	56	100
enzyme	136	0	2	13	99	91
gene	0	0	2	0	0	0
genus	15	0	0	9	100	63
location	11	0	3	10	79	52
measure	157	0	49	11	76	93
organism	59	0	26	23	69	72
organizatio	8	2	7	4	47	57
pathway	0	0	10	1	0	0
person	7	1	13	1	33	78
TOTALS	515	9	285	75	64	86

Table 2: Initial Named Entity results for EMPATHIE

the assistance of a molecular biologist, to allow an automatic evaluation of the PASTA terminology system using the MUC scoring software. Table 3 shows some preliminary results for the main terminology classes.

Name_Type	COR	INC	MIS	SPU	REC	PRE
protein	358	0	52	12	87	97
species	111	0	22	3	83	97
residue	175	0	4	13	98	93
site	53	0	34	10	61	84
region	19	0	24	0	44	100
2_struct	78	0	1	1	99	99
sup_struct	84	0	0	5	100	94
4_struct	115	0	5	3	96	97
chain	27	0	12	0	69	100
base	38	0	0	1	100	97
atom	42	0	2	10	95	81
non_protein	107	0	0	21	100	84
interaction	10	0	3	1	77	91
TOTALS	1217	0	159	80	88	94

Table 3: Initial Named Entity results for PASTA

5.2 Discussion

It should be stressed that these evaluation results are very preliminary, and we would expect them to improve substantially with further development.

The overall EMPATHIE template filling precision scores for both the development and evaluation sets are very close to the score of the LaSIE system in the MUC-7 evaluation (42%). Recall is noticeably lower however (47% in MUC-7), but this is certainly affected by the limited amount of training data available, giving a much smaller set of key words and phrases to use as cues for template fills. Also, it is clear that the EMPATHIE task requires much more specialist domain-specific knowledge than the MUC tasks, which typically require only general knowledge of companies and business procedures. The EMPATHIE task, as the process of manually filling the templates has demonstrated, can only be performed with the use of detailed domain knowledge, very little of which has been incorporated into the system. For example, a single mention of ‘cyanide’ in one of the evaluation texts causes

its entry as an ‘inhibitor’ in the manually filled template, though no explicit information in the text would allow it to be classified as such. Only domain-specific knowledge that cyanide is usually an inhibitor allows it to be classified in this case. Such cases are missed completely by the system because the specific knowledge required has not been entered, mainly due to the fact that the developer is not an expert in the domain.

Further consultation with experts would allow more domain-specific information to be entered, improving recall in particular. With this, and a more extensive training set, it should be entirely possible for system performance on the EMPATHIE task to equal the best MUC-7 scores (48% recall, 68% precision, from different systems).

The terminology recognition results are more encouraging, and compare favourably with MUC named entity results, particularly the PASTA results. It should be noted that both the EMPATHIE and PASTA terminology recognition tasks require the recognition of a considerably broader class of terms than the MUC named entity task and that considerably smaller sets of training data were available. The discrepancy between the EMPATHIE and PASTA results on this task can probably be explained by the fact that there was in fact no training data available specifically for the EMPATHIE task before the evaluation was carried out, only the informal feedback of biologists looking at system output. Furthermore, the annotation of texts for the EMPATHIE terminology task was carried out by a larger group of people than carried out the PASTA annotation task and without a formal annotation specification. Thus, this annotated data is almost certainly less consistently annotated and the results should therefore be interpreted with some caution.

6 Conclusion

Between these two projects much of the low-level work of moving IE systems into the new domain of molecular biology and the new text genre of journal papers has been carried out. We have generalised our software to cope with longer, multi-sectioned articles with embedded SGML; we have generalised tokenisation routines to cope with scientific nomenclature and terminology recognition procedures to deal with a broad range of molecular biological terminology. All of this work is reusable by any IE application in the area of molecular biology.

In addition we have made good progress in designing template elements, template relations, and scenario templates whose utility is attested by working molecular biologists and in adapting our IE software to fill these templates. Preliminary evaluations demonstrate the difficulty of the task, but results are encouraging, and the steps to take to improve performance straightforward. Thus, we are optimistic that IE techniques will deliver novel and effective ways for scientists to make use of the core literature which defines their disciplines.

Acknowledgements

EMPathIE is a 1.5 year research project in collaboration with, and funded by, GlaxoWellcome plc and Elsevier Science. The authors would like to thank Dr. Charlie Hodgman of GlaxoWellcome for supplying domain expertise and Elsevier for supplying electronic copy of relevant journals. PASTA is funded under the UK BBSRC-EPRSC BioInformatics Programme (BIF08754) and is a collaboration between the Departments of Computer Science, Information Studies and Molecular Biology and Biotechnology at the University of Sheffield. The authors would like to thank Dr. Peter Artymiuk of the University of Sheffield for supplying his expertise in protein structures.

References

- D. Appelt, J. Hobbs, J. Bear, D. Israel, M. Kameyama, A. Kehler, D. Martin, K. Myers, and M. Tyson. SRI International FASTUS system: MUC-6 Test Results and Analysis. In *Proceedings of the Sixth Message Understanding Conference (MUC-6) Def (1995)*, pages 237–248.
- F. Bernstein, T. Koetzle, G. Williams, E.J. Meyer, M. Brice, J. Rodgers, O. Kennard, M. Shimanouchi, and M. Tasumi. The protein data bank: A computer-based archival file formacromolecular structures. *Journal of Molecular Biology*, (112):535–542, 1977.
- J. Cowie and W. Lehnert. Information extraction. *Communications of the ACM*, 39(1):80–91, 1996.
- H. Cunningham, K. Humphreys, R. Gaizauskas, and Y. Wilks. Software Infrastructure for Natural Language Processing. In *Proceedings of the Fifth Conference on Applied Natural Language Processing (ANLP-97)*, pages 237–244, 1997. Available at <http://xxx.lanl.gov/ps/9702005>.
- Defense Advanced Research Projects Agency. *Proceedings of the Sixth Message Understanding Conference (MUC-6)*. Morgan Kaufmann, 1995.
- Defense Advanced Research Projects Agency. *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, 1998. Available at <http://www.saic.com>.
- K. Fukuda, T. Tsunoda, A. Tamura, and T. Takagi. Information extraction: Identifying protein names from biological papers. In *Proceedings of the Pacific Symposium on Biocomputing '98 (PSB'98)*, pages 707–718, Hawaii, 1998.
- R. Gaizauskas, T. Wakao, K. Humphreys, H. Cunningham, and Y. Wilks. Description of the LaSIE system as used for MUC-6. In *Proceedings of the Sixth Message Understanding Conference (MUC-6) Def (1995)*, pages 207–220.
- R. Gaizauskas and Y. Wilks. Information Extraction: Beyond Document Retrieval. *Journal of Documentation*, 54(1):70–105, 1998.
- J.R. Hobbs. Description of the TACITUS system as used for MUC-3. In *Proceedings of the Third Message Understanding Conference MUC-3*, pages 200–206. Morgan Kaufmann, 1991.
- J.R. Hobbs. The generic information extraction system. In *Proceedings of the Fifth Message Understanding Conference (MUC-5)*, pages 87–91. Morgan Kaufman, 1993.
- K. Humphreys, R. Gaizauskas, S. Azzam, C. Huyck, B. Mitchell, H. Cunningham, and Y. Wilks. Description of the LaSIE-II system as used for MUC-7. In *Proceedings of the Seventh Message Understanding Conference (MUC-7) Def (1998)*. Available at <http://www.saic.com>.
- T. Rindflesh, L. Tanabe, J. Weinstein, and L. Hunter. Edgar: Extraction of drugs, genes and relations from the biomedical literature. In *Proceedings of the Pacific Symposium on Biocomputing '2000 (PSB'2000)*, pages 517–528, Hawaii, 2000.
- E. Selkov, S. Basmanova, T. Gaasterland, I. Goryanin, Y. Gretchkni, N. Meltsev, V. Nenashev, R. Overbeek, E. Panyushkina, L. Pronevitch, E. Selkov, and I. Yunis. The metabolic pathway collection from EMP: The enzymes and metabolic pathways database. *Nucleic Acids Res.*, (24):26–28, 1996.
- J. Thomas, D. Milward, C. Ouzounis, S. Pulman, and M. Carroll. Automatic extraction of protein interactions from scientific abstracts. In *Proceedings of the Pacific Symposium on Biocomputing '2000 (PSB'2000)*, pages 541–551, Hawaii, 2000.