Two Applications of Information Extraction to Biological Science Journal Articles: Enzyme Interactions and Protein Structures

Kevin Humphreys, George Demetriou, Robert Gaizauskas
Department of Computer Science, University of Sheffield
Regent Court, Portobello Street
Sheffield S1 4DP UK

Information extraction technology, as defined and developed through the U.S. DARPA Message Understanding Conferences (MUCs), has proved successful at extracting information primarily from newswire texts and primarily in domains concerned with human activity. In this paper we consider the application of this technology to the extraction of information from scientific journal papers in the area of molecular biology. In particular, 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. Progress to date provides convincing grounds for believing that IE techniques will deliver novel and effective ways for scientists to make use of the core literature which defines their disciplines.

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** [1,2]. The prototypical IE tasks are those defined by the U.S. DARPA 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 [3,4]. 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 very limited way, is the construction of databases of scientific information from journal articles, for use by researchers in

molecular biology.[a] 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 (enyzmes, compounds) with attributes (names, concentrations) are related by participating in the event in particular 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.

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 describes preliminary results. While neither of these systems is yet complete, 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)[4] 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*).

[a] The only other application of IE techniques to texts in the biological sciences of which we are aware is the work of Fukada et al. on identifying protein names in MEDLINE abstracts[5].

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. The participating systems are then run and their results are automatically scored against the human-produced output or *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 9000) in the Protein Data Bank (PDB) [6] 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.

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

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.

3.1 EMPathIE

The Enzyme and Metabolic Pathways application is particularly suitable for investigating the use of IE techniques due to the availability of a manually constructed database for the same application. The EMP database[7] contains over 20,000 records of enzyme reactions, collected from journal articles published since 1964. This provides training data for use in the development of the EMPathIE system, and allows evaluation of the extent to which the system can automate the construction of EMP database records.

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> :=              <PATHWAY-1> :=
  NAME: isocitrate lyase     NAME: glyoxylate cycle
  EC_CODE: 4.1.3.1           INTERACTION: <INTERACTION-1>

<ORGANISM-1> :=            <INTERACTION-1> :=
  NAME: Haloferax volcanii   ENZYME: <ENZYME-1>
  STRAIN: ATCC 29605         PARTICIPANTS: <PARTICIPANT-1>
  GENUS: halophilic Archaea                <PARTICIPANT-2>

<COMPOUND-1> :=            <PARTICIPANT-1> :=
  NAME: phenylhydrazone      COMPOUND: <COMPOUND-1>
                             TYPE: Product
<COMPOUND-2> :=            TEMPERATURE: 35C
  NAME: KCl
                           <PARTICIPANT-2> :=
<SOURCE-1> :=              COMPOUND: <COMPOUND-2>
  ENZYME: <ENZYME-1>       TYPE: Activator
  ORGANISM: <ORGANISM-1>   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-str-1997-5-2-1>:=
        RESIDUE_TYPE:  SERINE
        RESIDUE_NO:    "87"
        IN_PROTEIN:    <PROTEIN-str-1997-5-2-1>
        SITE/FUNCTION: "active site", "catalytic",
                       "interfacial activation", "calcium-binding site"
        SECOND_STRUCT: alpha-helix
        REGION:        'lid'
        ARTICLE:       <ARTICLE-str-1997-5-2-1>

<PROTEIN-str-1997-5-2-1>:=
        NAME:       "Triacylglycerol lipase"
        SCOP_CLASS: "Lipase"
        PDB_CODE:   1LGY
        IN_SPECIES: <SPECIES-str-1997-5-2-1>

<SPECIES-str-1997-5-2-1>:=
        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.

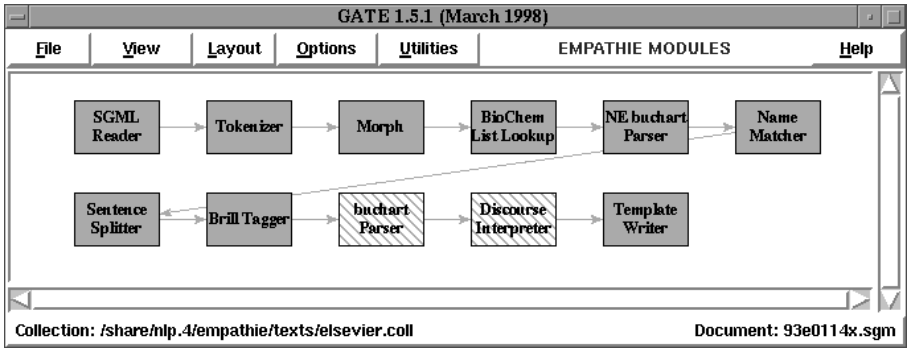

Figure 2: EMPathIE system modules within GATE

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 [8,9]. 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 [10]. LaSIE is neither as 'deep' as some earlier IE systems that attempted full syntactic, semantic and discourse processing [11] nor as 'shallow' as some recent systems that use finite state pattern matching techniques to map directly from source texts to target templates [12].

The processing modules which make up the EMPathIE system are shown in figure 2, within the GATE development environment [13]. The PASTA system is similar and reuses several modules, within the same environment. 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 excluded 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, assembled 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.

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.

An initial domain model for the EMPathIE metabolic pathway task has been manually constructed, directly from the template definition, and subsequent refinement of the model 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

Currently, a complete prototype EMPathIE system exists which can produce filled templates as specified above. The terminology recognition portion of the system has been informally reviewed by molecular biologists who have found its performance to be remarkably good, and resources for evaluation of the IE results are being manually constructed.

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 the assistance of a molecular biologist, to allow an automatic evaluation of the PASTA terminology system using the MUC scoring software.

Table 1 shows some preliminary results for the main terminology classes. The columns show: the number of responses proposed by the system (POSsible), the number of items in the texts of the given class (ACTual), the number of items the system correctly identified (CORrect), the number the system missed (MISsing), the number the system spuriously proposed (SPURious) and the standard metrics of RECall and PREcision, discussed in section 2 above. It should be stressed that these results are very preliminary, and we would expect them to improve substantially with further development. However, they compare favourably with MUC named entity results.

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.

Name_Type	POS	ACT	COR	MIS	SPU	REC	PRE
protein	410	370	358	52	12	87	97
species	133	114	111	22	3	83	97
residue	179	188	175	4	13	98	93
site	87	63	53	34	10	61	84
region	43	19	19	24	0	44	100
second_struct	79	79	78	1	1	99	99
super_second_struct	84	89	84	0	5	100	94
quatern_struct	120	118	115	5	3	96	97
chain	39	27	27	12	0	69	100
base	38	39	38	0	1	100	97
atom	44	52	42	2	10	95	81
non_protein	107	128	107	0	21	100	84
interaction	13	11	10	3	1	77	91
TOTALS	1376	1297	1217	159	80	88	94

Table 1: Initial Named Entity results for the PASTA system

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. 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

1. J. Cowie and W. Lehnert. Information extraction. *Communications of the ACM*, 39(1):80–91, 1996.

2. R. Gaizauskas and Y. Wilks. Information Extraction: Beyond Document Retrieval. *Journal of Documentation*, 54(1):70–105, 1998.
3. Defense Advanced Research Projects Agency. *Proceedings of the Sixth Message Understanding Conference (MUC-6)*. Morgan Kaufmann, 1995.
4. Defense Advanced Research Projects Agency. *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, 1998. Available at `http://www.saic.com`.
5. 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, January 1998.
6. 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.
7. 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.
8. R. Gaizauskas, T. Wakao, K Humphreys, H. Cunningham, and Y. Wilks. Description of the LaSIE system as used for MUC-6. In[3], pages 207–220.
9. 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[4].
10. J.R. Hobbs. The generic information extraction system. In *Proceedings of the Fifth Message Understanding Conference (MUC-5)*, pages 87–91. Morgan Kaufman, 1993.
11. 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.
12. 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[3], pages 237–248.
13. 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, March 1997.