Information Extraction from Biological Science Journal Articles*

Kevin Humphreys and **George Demetriou** and **Robert Gaizauskas**
Department of Computer Science, University of Sheffield
Regent Court, 211 Portobello Street, Sheffield S1 4DP UK

1 Introduction

Information extraction technology, as defined and developed through the U.S. DARPA Message Understanding Conferences (MUCs) [DARPA, 1998], has proved successful at extracting information primarily from newswire texts and 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.

Information extraction (IE) may be defined as the activity of extracting information about 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]. Prototypical IE tasks are those defined by the MUC evaluations which have involved analysing newswire text to fill templates for scenarios such as management succession events, or rocket launchings. Performance of this technology is still not at human levels for all tasks, but is approaching human levels for some tasks (e.g. the recognition and classification of named entities in text) and is at a level for other tasks at which comparable technologies, such as information retrieval and machine translation, have found useful application.

One subject area and text genre to which IE techniques have not yet been applied is the extraction of information from scientific journal articles for use by molecular biologists. In this paper we describe the IE system we developed to take part in the DARPA MUC evaluation exercises, how we have modified it for bioinformatics applications, and two specific bioinformatics 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 journal papers.

2 The LaSIE System

The Large Scale Information Extraction (LaSIE) system has been designed as a general purpose IE system which can conform to the MUC-7 task specifications for named entity identification, coreference resolution, IE template element and template relation filling, and the construction of scenario-specific IE templates. The precise specifications of these tasks may be found in DARPA [1998].

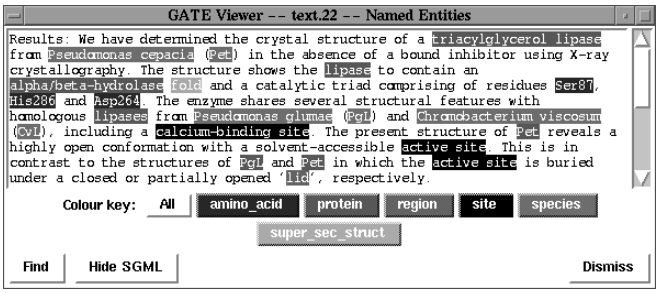
LaSIE has a pipeline architecture which processes a text one sentence at a time and consists of three principal processing stages. *Lexical preprocessing* reads and tokenises the raw input text, performs phrasal matching against lists of proper names, identifies sentence boundaries, tags the tokens with parts-of-speech, performs morphological analysis. *Parsing and semantic interpretation* builds lexical and phrasal chart edges in a feature-based formalism then does two pass chart parsing, pass one with a special named entity grammar, pass two with a general grammar, and, after selecting a 'best parse', which may have only partial coverage, constructs a predicate-argument representation of each sentence. *Discourse interpretation* adds the information from the predicate-argument representation to a hierarchically structured semantic net which encodes the system's world and domain model, adds additional information presupposed by the input, performs coreference resolution between new and existing instances in the world model, and adds any information consequent upon the new input. The output of discourse interpretation is a discourse model which can be viewed as a specialisation of the domain model for the text at hand. The final templates the system produces are read off this model. See Gaizauskas and Humphreys [1997] for more details.

3 Bioinformatics Applications

We are currently investigating the use of the approach to IE represented by the LaSIE system in two separate bioinformatics research projects. The Enzyme and Metabolic Pathways Information Extraction (EMPathIE) project aims to investigate the use of the LaSIE approach for the extraction of details of enzyme reactions from the SGML source of articles from the journals Biochimica et Biophysica Acta and FEMS Letters. 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 scientific journal abstracts and full articles.

*Two research projects are described in this paper. The EMPathIE project is funded by GlaxoWellcome plc and Elsevier Science. The PASTA project is funded under the UK BBSRC/EPRSC BioInformatics Programme (50/BIF08754).

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 adaptation for domain-specific terminology. This has resulted in an independent terminology identification subsystem, prior to any general syntactic analysis. 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. This contrasts with the original LaSIE system where the main sources were: case-sensitive terminology lexicons; part-of-speech tags, which reflected the use of capitalisation; and hand-constructed 'named entity' grammar rules. Results of the current terminology subsystem in the PASTA domain are shown below:



Besides the terminology in the biochemical domain, a further significant difference from newswire applications is the structure of the texts. Scientific articles typically have a rigid structure, including abstract, introduction, method and materials, results, and discussion sections. For particular applications some sections can be targeted for detailed analysis while others can be skipped entirely.

EMPathIE An initial domain model for the metabolic pathway task has been manually constructed, directly from a MUC-style IE template definition. This has involved 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 were also added for each of the template slots, `concentration`, `temperature`, etc., and rules were then 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 the domain model involves extending the semantic network subhierarchies and the addition of coreference constraints on the hypothesised instances, based on available training data.

The template below 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 `phenylhydrazone`, which has the role of *product* of the interaction at a temperature of 35C. The second is the compound *KCl*, which has the role of *activator* at a concentration of 1.75M.

```
<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: <INT.-1>
<INT.-1> :=
  ENZYME: <ENZYME-1>
  PARTICIPANTS: <PART.-1>
                <PART.-2>
<PART.-1> :=
  COMPOUND: <COMPOUND-1>
  TYPE: Product
  TEMPERATURE: 35C
<PART.-2> :=
  COMPOUND: <COMPOUND-2>
  TYPE: Activator
  CONCENTRATION: 1.75 M
```

PASTA Template design for the PASTA project has led to the definition of four entity templates and two relational templates. The entity templates are for residues (residue type, number, site/function, protein and structure type in which found), proteins (name, family, species in which found), species (name) and text source (journal, title, author).The relational templates aim to capture relations of structural equivalence (between one molecule and a different molecule when a protein shares similar structural characteristics with another protein in terms of the residues involved) and three-dimensional arrangement (between residues within one structure).

4 Conclusion

At this stage an end-to-end prototype EMPathIE system exists which can produce filled templates as specified above. A sizeable test collection is being assembled and in conjunction with the EMP database will shortly be used for a quantitative evaluation of the entire system. The terminology recognition portion of the system has been informally reviewed by molecular biologists who have found its performance to be remarkably good. The PASTA system has been implemented as far as the terminology recognition stage and a corpus of 1500 abstracts and 600 full journal papers is being assembled and partially manually annotated for purposes of development and evaluation. 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. For a more detailed description of these projects see `http://www.dcs.shef.ac.uk/research/groups/nlp/`.

References

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

[DARPA, 1998] Defense Advanced Research Projects Agency. *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, 1998. `http://www.saic.com`.

[Gaizauskas and Humphreys, 1997] R. Gaizauskas and K. Humphreys. Using a semantic network for information extraction. *Journal of Natural Language Engineering*, 3(2/3):147–169, 1997.