

BiRD: An Automatic System to Build Resource Databases for Researchers, ESRC Grant number RES-000-23-0010, September 2003-September 2006: Final Report

1. Background

In the field of Natural Language Processing (NLP), the specialist community has accumulated a large amount of NLP resources that it has developed over many years including software (part-of-speech taggers, parsers, various corpus analysis tools) and data (evaluation corpora and datasets, lexicons, gazetteers, terminology databases). Most of these resources are freely available, which helps other researchers to save effort on developing them and allows for direct comparisons of experimental results with previous work. Unfortunately, finding these resources on the web is not straightforward. Traditional keyword search is often too costly in terms of time and effort. While it is possible to consult collections of web links on a research topic, these collections have limited coverage and quickly fall out of date, as they are manually compiled and maintained. It is clear that existing general-purpose tools are not sufficient for researchers and specifically designed search tools should be used instead. The purpose of this project is to build a system that enables researchers to easily locate the information they require. It is expected that the implemented system will prove useful for researchers, as well as students and educators, industrial companies and government bodies. The first implementation of the system is tuned to the field of computational linguistics but we are confident that this will not restrict the generality of the system, and the architecture will be applicable to any domain.

The automatic acquisition of knowledge from texts is not a new topic in computational linguistics. Text mining techniques have been used to address such diverse practical tasks as the discovery of previously unknown information about migraine headaches from within the medical literature (Swanson and Smalheiser, 1997), and study of the impact of publicly financed research on industrial advances (Narin et. al., 1997). Web crawling and extraction techniques have been used to automatically build the publicly available Autonomous Citation Indexing system, *CiteSeer*, (Bollacker, 2000) which finds scientific papers on the Internet and interlinks them based on their references.

It is thus clear that the BiRD project does not address completely unexplored problems. However, both the way in which different methods have been combined and the end product itself are novel. For example, in the MUC competitions, participants were asked to complete templates using information from a single document. In our project, we acknowledge the fact that a researcher may wish to seek further information about the entities related to a particular resource. This type of information demand is met by applying techniques from multi-document summarisation and cross-document coreference resolution. Furthermore, the systems competing in the MUC IE evaluations processed documents with a standardised format. In the current project, extremely heterogeneous documents are used as the basis for information IE. Finally, while systems such as *CiteSeer* do not use linguistic information to build their databases, in the BiRD project, linguistic information is the main resource used to build and maintain the database.

The fields of multi-document summarisation and cross-document coreference are in the early stages of development. To the best of our knowledge, there are no coreferentially annotated corpora available which indicate the relations between entities in different documents (Mitkov, forthcoming). The corpus developed in this project has been annotated for cross-document coreference. As a text genre, emails have been an under-researched field of discourse analysis, but in recent years, due to an increase in their use, interest in emails has increased rapidly, inspiring corpus builders to include them in general-purpose corpora (e.g. the American National Corpus (Ide & Suderman, 2004)). To the best of our knowledge there has been little investigation of the discourse structure of web pages and research that can be directly applied to computational and corpus linguistics is particularly sparse. The investigation of the structure of web pages and the discourse of resource announcements has been essential in the BiRD project.

2. Objectives

The overall aim of the project was to develop a fully-automatic software agent to mine email messages, and pages linked to them, for information announcing the availability of new resources in a research field, and populate a database with this information. The system implemented here was designed to extract information from documents in the field of computational linguistics. It was decided that specialised mailing lists such as *Corpora* and *Elsnet* would serve as sources of email messages containing a relatively high proportion of relevant documents.

From the outset, it was envisaged that the aim of the BiRD project would be met by addressing the following primary objectives:

1. Development of a software agent that will mine email messages and linked web pages for information announcing the availability of new resources in a research field. The agent will update a user's database in order to incorporate the extracted information.
2. Contribution to the fields of linguistic theory and information extraction, particularly with regard to multi-document coreference, named entity recognition in the research domain (exploiting a new typology of named entities), multi-document summarisation, and template filling/merging.
3. The creation and annotation of a 250,000 word corpus of HTML pages with entities of interest and the relations holding between those entities from an inter- and intra-document perspective (within document and cross-document coreference).
4. An examination of the discourse structure of web pages and email messages.
5. Semi-automatic acquisition of templates
6. Contribution to knowledge and understanding of the relationships that hold between entities within documents and between documents
7. Evaluation in the fields of cross-document coreference resolution and multi-document summarisation
8. Propose annotation schema and guidelines

At the initial stages of the project, a pilot study of domain documents was carried out. As a result of the study, various assumptions were dropped and a more accurate view on the context of our research was derived. Firstly, it was noted that email is typically a poor source of information with regard to the announcement of the availability of resources such as corpora and NLP software. Instead, we recognised that web pages are more suitable sources. This motivated a change of emphasis in the project away from announcements delivered by means of email messages and toward the processing of heterogeneous web pages. At the same time we found that the processing of heterogeneous web pages is dependent upon the adequate treatment of a range of issues such as highly variable vocabulary and discourse structure of the web pages and presence of relevant information in both unstructured and semi-structured text. Furthermore, given that this type of document is not posted directly to domain-specific mailing lists, the decision was made to implement software capable of automatically locating web pages presenting NLP resources.

In light of this updated knowledge, the tasks of the project were formulated as follows.

Implementation of existing methods. In many cases, research problems have been addressed by pre-existing methods. Following a literature review, the best-suited methods were chosen for implementation as modules within the BiRD information facility. These include existing methods for text categorisation, information extraction (IE), language identification, coreference resolution, multi-document summarisation (MDS), and named entity recognition (NER). Their implementation serves as a solid foundation for optimisation, domain tuning and improvement of the system by novel methods. Some existing software, such as the GATE NER component, could be easily plugged in to meet some generic system requirements whereas others required considerable domain tuning. Two examples of the latter are the customisation of Rainbow, the text categorisation toolkit, to operate on a novel representation of documents, and the semi-automatic extension of the COLLATE ontology with terms extracted from a domain corpus.

Development of new methodologies. Despite the existence of several practical systems for IE that have been developed since the 1990s when a definition for IE was accepted and formalised as a result of the MUC conferences, these were considered unsuitable because they are geared toward scenarios very different from the scenario addressed by the BiRD project. Several new methodologies have been developed throughout the project in order to address this finding. They comprise new approaches to:

1. Text categorisation geared toward the subsequent application of information extraction methods,
2. Extension of the double classification approach to IE for the current task, and
3. Multi-document summarisation incorporating the web-based acquisition of background knowledge and a probabilistic approach to generation of references to entities.

Development of evaluation data. Given that one of the requirements of the BiRD project is the development of evaluation data for use in training ML methods for IE and text categorisation, and to make the resultant datasets publicly available, the following methodology was adopted. Firstly, pre-existing hubs and the pre-classified email messages stored on mailing list archives such as *Elsnet*, were used to collect relevant documents for manual annotation. A range of freely available annotation tools was evaluated and PALinkA (Orasan, 2003) was selected due to the ease with which the annotation schema that it encodes can be customised. The collected documents were then manually annotated. 250 documents were annotated with respect to the appropriate IE templates, while more than 700 were labelled for document categories. The

evaluation of the multi-document summarisation method was facilitated by the manual construction of evaluation templates capturing an exhaustive set of relevant information about various domain concepts.

Experimental study of the methods: On the basis of the developed evaluation data, we carried out an experimental evaluation of separate system components (text categorisation, named entity recognition, information extraction, and multi-document summarisation) as well as a user evaluation of the system was carried out.

System integration: Both the existing and newly developed methods to address the goals of the project have been implemented in specific processing modules that have been integrated into a unified system (see Figure 1 for an overview of the system). After a web page is supplied to the system, it is sequentially passed through the processing units, with each unit placing XML mark-up into the page. The procedure is pipelined in such a way that each unit benefits from the annotations produced by the previous one. At the end of the process, the page receives normalised structural, linguistic, and semantic mark-up, which are later used for training and evaluation of the IE unit.

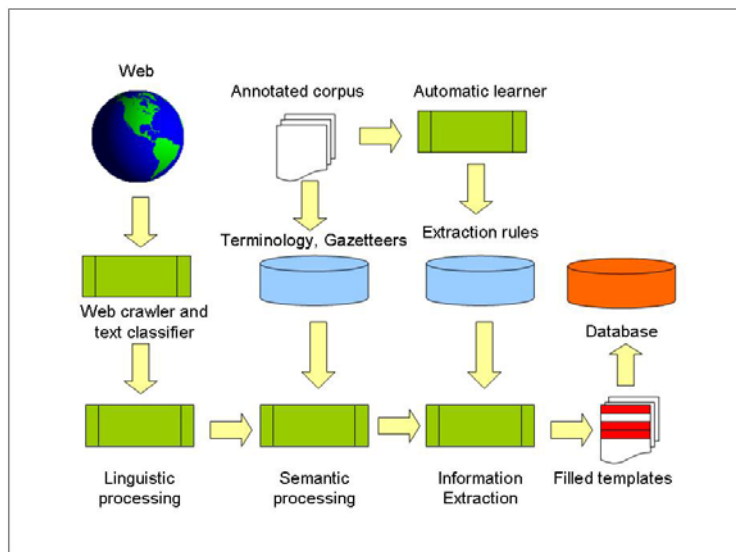


Figure 1. Overview of the BiRD system.

3. Methods

The choice of the methods to be implemented in various stages of document processing was driven primarily by the consideration of performance (effectiveness and efficiency) in the application domain at hand. At the same time, the methodological design of the system was chosen to maximise its capacity to be ported to new application domains. This section outlines the methods used by the major processing units of the BiRD system: domain crawler, text filtering, terminology and named entity (NE) recognition, intra- and cross-document NP co-reference resolution, information extraction (IE) and multi-document summarisation (MDS).

3.1 Domain crawler

To monitor the web for relevant pages continuously, a focused crawling methodology has been implemented, which operates based on the assumption the best route for the discovery of pages on a very narrow topic is to consult web pages with collections of web links (the so-called “hubs”), rather than key word search. The methodology is based on the HITS algorithm for link analysis (Kleinberg 1999), which, starting from a few seed hubs, seeks out more hubs using back link queries¹, and authoritative pages that are being referred to by such hubs. The algorithm performs these steps iteratively, at each iteration re-computing relevance scores for hubs and pages, whereby relevance is operationalised via the number incoming and outgoing links to and from pages known to be relevant. In this way, the algorithm enlarges both sets of URLs and terminates after a certain stopping criterion has been reached.

¹ Our specific implementation makes use of an API to the Yahoo! search engine.

3.2 Text filtering

To identify relevant documents among those retrieved by the crawler, the text filtering component makes use of machine learning methods for text categorisation². An evaluation corpus was used in extensive experiments with a variety of existing text categorisation methods. In addition, a new technique to derive web page vectors has been developed (Pekar, Evans, Mitkov, 2004), which aims to improve classification of pages relative to an information extraction template. The procedure exploits unsupervised IE techniques to give weight to terms that are likely to be either highly indicative cues for identifying template fillers or be template fillers themselves, and downplay terms that have little relevance to the template. In this way, the method attempts to pinpoint similarities between pages that are characterised by extremely varying vocabularies, but which nonetheless can be relevant to the same IE template. Our evaluation indicated that the developed method allows the creation of much more efficient document representations, and, in a number of cases, also to enhance the classification accuracy.

3.3 Terminology and NE Recognition

The system component for terminology and NE recognition relies on a domain gazetteer and terminology repository acquired automatically from a domain corpus. The acquisition of the lexical material is carried out by means of a technique similar to those developed by Juteson and Katz (1996) that searches statistically significant word sequences in the corpus matching pre-defined patterns of part-of-speech (PoS) tags indicative of domain terms, such as *(Adj|Noun)*Noun*. The unit also performs pattern-matching to locate acronyms and abbreviations and uses them to delineate multi-word terms and identify their variants. In order to recognise the semantic types of the terms, finite-state transducers are used that fire on certain constituents of the terms, their orthography and PoS tags.

Recognition of terminology and NEs is performed in two steps. In the first step, common named entities (person names, locations, and dates) are recognised with the help of the pattern-based approach implemented in the ANNIE system (Cunningham, Maynard, Bontcheva, & Tablan, 2002). Domain-specific terms and NEs are recognised using domain gazetteers and terminology lists.

The first stage of NE recognition produces NE and terminology tags with high precision, but has poor coverage, especially in the case of domain-specific items, which are often missing from the gazetteers. In the second stage, we make use of the already available NE annotations and the layout cues to leverage mark-up of NEs and terms that remain unmarked to this point.

For this purpose two methods are employed, both of which are general-purpose. The first one is a list extraction technique akin to the one used in (Etzioni, Cafarella, Downey, Kok, Popescu, Shaked, Soderland, Weld & Yates 2004). It consists of (i) locating itemisation lists in the page that contain a sufficient number of marked-up NEs and terms, the NEs are used as seed data (see Figure 2); (ii) inducing a list-specific ‘wrapper’ from the seeds; and (iii) applying the wrapper to mark up remaining terms and NEs in the list.

```
Candace Kamm, <ORGANIZATION>AT&D</ORGANIZATION>  
Lin-Shan Lee, <ORGANIZATION>Taiwan University</ORGANIZATION>  
<PERSON>Susann Luperfoy</PERSON>, Akamai Technologies  
<PERSON>Patti Price</PERSON>, SRI International  
<PERSON>Owen Rambow</PERSON>, <ORGANIZATION>AT&D</ORGANIZATION>
```

Figure 2. An itemisation list with seed named entities marked-up.

The other technique is an extension of the previous one, to operate on grammatical text. It finds sentences which contain enumerated expressions, each of which is separated by the same punctuation, and possibly a conjunction. Assuming that enumerations comprise expressions with similar semantics, all of the expressions delimitable by means of the commas and conjunctions are marked-up by the same tag as the seeds.

3.4 Intra- and cross-document NP coreference resolution

In order to detect coreference chains involving NEs of interest, we used the light-weight techniques for NE coreference resolution introduced in (Bontcheva, Dimitrov, Maynard, et al. 2002), which detect coreference

² The implementation of these methods was obtained from the freely available Rainbow toolkit (McCallum, 1996).

based on orthographic cues, such as those capturing different variants of a person's name (e.g. *John Brown*, *J.Brown* and *Mr. Brown*) and regular ways to create an acronym for an organisation (e.g. *UN* for *United Nations*). Common noun phrases are introduced into coreferential chains using the terminology database containing semantic labels for each term: noun phrases that appear in the close vicinity of a term and that match its semantic label are taken to be co-referent with the term.

For each NE recognised in the document, a set of its variants is generated automatically (the case of named entities) or retrieved from the terminological database (in the case of abbreviations and acronyms). These variants are used to first create coreference chains inside each document. Then document-specific chains are compared across documents and united into cross-document chains. Prior to feeding the document to the IE module, all entities in a coreferential chain are substituted for the normalised variant of the entity.

3.5 Information Extraction

The specific characteristic of the IE task at hand can be described as the one-template-per-document IE scenario: given a set of documents that describe a particular type of entity or event (such as conference announcements), the goal is to fill an IE template from an entire document.

Within the BiRD project, two different IE methods to address this goal have been experimentally studied. The first one, the double classification method (De Sitter & Daelemans, 2003), is inspired by the observation that when humans need to extract some facts from a document, they scan it quickly and only read the parts that look most relevant. In a similar manner, the double classification method uses the first classifier to identify document fragments that are likely to contain template fillers. The second classifier classifies tokens inside the promising fragments to more precisely pinpoint the filler. Within the BiRD project, we investigated a number of techniques to optimise the difficulty of each of the two classification sub-problems, in order to increase the overall accuracy of the method, and proposed a method that post-processes the output of the double classification approach to delimit each token or sequence of tokens in a document as the most likely fillers for the template.

The second IE method investigated in the project relies on manually constructed information extraction patterns. While limiting the domain portability of the system, this approach appears justified by the purposes of the current project whose aim is to deliver an accurate IE system in a single application domain. To facilitate construction of IE patterns, knowledge engineers were supplied with a corpus with manually annotated documents and a concordancer.

3.6 Multi-Document Summarisation

An entity-based approach to MDS was implemented. It exploits the templates produced by the IE process, incorporating techniques to allow variation in the generation of references to entities. The MDS method additionally relies on queries to internet search engines which are used as a means to obtain background knowledge that is not present in the source documents alone. The method has been investigated in Evans (2006, submitted).

4. Results

This section describes the results of experimental work on the project, which comprises the evaluation of existing and newly developed methods implemented in separate system components and the evaluation of the entire system from the user perspective.

4.1 System components

4.1.1 Text categorisation

Within the text categorisation task, we investigated various machine learning algorithms on a corpus of documents representative of the knowledge domain at hand (reported in Pekar, Evans, Mitkov, 2004). In these experiments we examined two methods to construct document vectors necessary for their automatic classification: the traditional “bag-of-words” representation and a newly proposed representation that aims to improve classification relative to an information extraction template.

The results of the evaluation indicate that the use of primary IE patterns to represent the content of the documents is preferable when the documents need to be assigned to categories relevant to the IE task. We first found that the quality of categorisations did not deteriorate when using this method to represent

documents while the term space was reduced to around half the size of the baseline method. Secondly, in certain cases we found a statistically significant improvement in categorisation accuracy, in particular for categorisation of web pages, which, unlike email messages, come from extremely diverse sources and are often characterised by very diverse vocabularies. The use of primary IE patterns thus seems to emphasise the ability of diverse web pages to be used for filling-in a particular IE template.

4.1.2 Enrichment of documents with semantic annotations

We conducted a study of the effect the different document pre-processing techniques have on the quality of information extraction (Pekar 2005; Pekar & Evans 2006, submitted). These techniques are meant to enrich documents with more semantic information on tokens and relations between them that will eventually be useful for automatic learning of IE rules. In these experiments, we looked at the performance of the double classification IE method when (i) neither layout nor NE and terminology recognition is performed; (ii) when the documents are enriched with layout tags and (iii) when NE and terminology annotations are also added to the document. The evaluation has shown that introduction of additional layout tags into the documents leads to an increase in the quality of IE comparable to the use of terminology and NE tags (in both cases the average improvement across template fields was up to 3%). This finding appears to indicate that while terminology and NE tags provide additional semantic evidence about tokens in the document, its layout contains useful evidence about the relationships between tokens.

4.1.3 Information Extraction

The IE component of the system implements the double classification approach to IE. We conducted an investigation into several parameters of the method in order to optimise its two classification subproblems and eventually improve its overall performance (reported in Pekar & Evans, 2005; Pekar & Evans 2006, submitted). These experiments have shown that finding a balance of factors influencing the distribution of the task difficulty between the two classifiers helps to improve the overall IE performance. In particular, doing so improved F-measure by 26% in comparison with the original configuration of the method described in (De Sitter & Daelemans 2003).

The double classification method aims to extract all document tokens instantiating template fields, which is a very difficult and error-prone task. However, what is often needed instead is accurate extraction of a single filler which may consist of a single token or a sequence of tokens. We have developed a new method for the identification of such fillers in the output of the IE method. This technique takes advantage of the evidence for the best filler in the form of the relative position of tokens labelled as positive by the second classifier, the frequency of the token sequences, and the frequency of their component tokens. The evaluation has shown that the method coupled with the double classification approach performs consistently better than extraction rules constructed manually.

4.1.4 Multi-Document Summarisation

The experimental module for MDS was evaluated by first running it over a corpus of 1571 email messages sent to the *Corpora* list. Evaluation templates were produced by hand and the information produced in the output of the MDS module was compared with them. Due to the extremely arduous and time-consuming nature of manual template production in an unrestricted multi-document scenario, just two templates were manually produced. The implemented MDS module was assessed over a range of criteria. From the perspective of their information content, the summaries obtained scores for precision of 0.72 and 0.78. From the perspective of the validity of the generated background knowledge, the MDS module obtained an accuracy of 0.52 and 0.57 in the output summaries. Additional details on the evaluation of this module are provided in Evans (2006, submitted).

4.2 User evaluation of the system

To perform user evaluation of the entire BiRD system, we designed a questionnaire on various existing NLP resources and previous (since 1996) or forthcoming conferences on NLP. The questions were constructed about non-empty, randomly selected fields in the certain pre-specified columns of the BiRD database. The questionnaire included 10 “specific fact” questions, i.e. questions to which answers can be retrieved from a single document (e.g., “*Under which platform(s) can TreeTagger be run?*” and “*Is the GATE system freely available for download?*”), and 5 “generalisation” questions that required the user to construct answers from multiple documents (e.g., “*What PoS taggers for German are available for free?*” and “*Which written corpora exist for Spanish?*”).

Four users were asked to answer these questions using the BiRD system and a search engine of their choice (the design ensured that each user is not presented with the same question twice). For each of the search facility used, time to retrieve the answers, the number of correct and incorrect answers, and the number of missing answers were measured. After the evaluation session the users were asked to share their comments on the usability of the interface of the systems used.

	Time per session	Wrong answers	Missing answers
BiRD	19.25	1.68	1.75
Internet	35.25	1.25	0.75

Table 1. User evaluation of the BiRD system.

As one can see, the use of the BiRD system results in only slightly bigger number of incorrect and missing answers, but requires around 2 times less time than popular search engines to retrieve an answer.

5. Activities

Project participants published the results of the BiRD project at six international conferences in Europe (IMIR'03, LREC'04, RANLP'05, NLDB'05), North America (ALC/ACH'05), and Asia (ILCP'05). A collaborative meeting with the participants of the COLLATE project was held in DFKI, Saarbruecken, Germany, which similarly aims to provide an intelligent information facility in the field of NLP. The contacts with DFKI resulted in exchange of valuable resources between the two projects (e.g., BiRD has employed the COLLATE ontology as the core for its terminological database). The results of the projects were additionally disseminated at meetings and open-day sessions for industrial visitors at the research group (AskJeeves, Amdocs, QuestionMark, Translution, R&D Dpt. of the New Cross Hospital, E-Know, E-Trad), colleagues from institutions in the UK and around the world (University of Manchester, University of Edinburgh, Oxford University, University of Barcelona, University of Saarland, University of Tuebingen), as well as staff and students at the University of Wolverhampton.

6. Outputs

The major outputs of the BiRD project include:

(1) Scientific publication. Five conference papers have been presented at scientific conferences relevant to the topic of the project. Two articles have been submitted to two relevant journals ("Applied Artificial Intelligence" and "Language Resources and Evaluation"). A keynote speech describing the information extraction methodology used in the project has been delivered at ILCIP'05. These publications are additionally being disseminated via the project website³.

(2) A corpus of web pages devoted to NLP resources and conferences has been prepared for dissemination. Its size is 500 documents (600,000 words), comprising 250 documents annotated manually and 250 documents annotated automatically for IE templates; all 500 documents have automatic annotation for named entities, domain terminology, intra- and cross-document NP co-reference. For copyright reasons, the corpus contains only standoff annotations and URLs of the full-text documents as well as their permanent URLs at the internet archive web site. A fragment of the corpus is shown in Annex 3. The corpus is currently in the process of being submitted to the ESRC Data Archive.

(3) An on-line search facility that enables access to the BiRD database⁴.

7. Impacts

Several meetings with the head of the Research and Development Directorate at the Royal Wolverhampton Hospitals NHS Trust, led to an expression of interest in tuning the BiRD system for use in the medical domain. The system was regarded as being of particular interest to medical research staff in the context of their continued assessment.

8. Future Research Priorities

³ <http://clg.wlv.ac.uk/projects/BiRD/>

⁴ Access to the conferences database is at <http://clg.wlv.ac.uk/BiRD/conf/>, to the NLP resources database is at <http://clg.wlv.ac.uk/BiRD/soft/>.

There are several areas in which additional research would yield further progress in the BiRD project. With regard to issues relevant to the original statement of the proposal, we note that the implemented system consists of a prototype whose efficiency is not currently optimal. It would be beneficial to improve the format of the database with a view to improving its efficiency. The exploitation of the database derived by the BiRD information extraction system could also be extended. At present, the summaries generated by the multi-document summarisation system for person entities and organisation entities are far superior to those generated for other types of entity. The tuning of multi-document summarisation to cater for additional types of entity would be beneficial. As the results of the user evaluation show, the user-friendliness of the search interface could be improved by enhancing its ability to deal with the variability of queries. The automatic generation of RSS feeds on the basis of the evolving database would be another beneficial exploitation of resources developed during the project. Finally, the continued maintenance and longitudinal evaluation of the BiRD system will enable the project to make a long-term impact on researchers both in the field of computational linguistics and in other disciplines.

Beyond the scope of the initial project proposal, promising lines of research include the analysis of social networks that have arisen in particular research fields. Such analysis may allow for the identification of particular schools of thought that emerge within a field, authoritative figures associated with those schools, as well as the emergence of new trends in the research being carried out. Visualisation of the research activity within a field is another research line that may profitably be pursued. It will be useful to continue to explore the commercial exploitation potential of the approach developed during the BiRD project. Possible applications include the processing of document collections from different genres and domains that share a communicative goal, but that manifest a heterogeneous manner of presentation (e.g. collections of university prospectuses, archives of job announcements and CVs, or financial bulletins). Useful applications will depend upon the extension of the BiRD system to new disciplines. Such extensions will be based, in the first instance, upon the replacement of specific rule-based components with generically effective machine learning algorithms.

ANNEX 1. BiRD publications

1. Evans R. 2004. *Building the Corpus Used in the BiRD Project*. Technical Report. CLG. University of Wolverhampton.
2. Orasan, C., Evans, R. and Mitkov, R. 2003. An automatic system to build resource databases for researchers. *Proceedings of the Information Mining and Information Retrieval*, 770-774. Crete, Greece.
3. Pekar V., Evans R., Mitkov, R. 2004. Categorizing web pages as a pre-processing step for information extraction. In *Proceedings of LREC-2004*. Lisbon, Portugal. pp.723-727.
4. Pekar V. 2005. Information Extraction from email announcements. In *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB-05)*. Alicante, Spain.
5. Pekar V. and Evans R. 2005. Automatic discovery of NLP resources on the web. In *Proceedings of ACH/ALLC'05*. Victoria, Canada.
6. Pekar V. 2005. Information extraction from heterogeneous documents. In *Proceedings of the 2nd Seminar of Languages, Cognition and Information Processing*. University of Chongqing, China.
7. Pekar V. and Evans R. 2005. Optimizing Classification Problems for the Double Classification Method for Information Extraction. In: *The International Conference on Recent Advances in Natural Language Processing (RANLP-05)*.
8. Pekar V. and Evans R. An Integrated Approach to Information Extraction from Heterogeneous Web Pages. In *Journal for Applied Artificial Intelligence*. (submitted).
9. Evans R. Entity-Based Summarization of Email Archives. In *Language Resources and Evaluation*. (submitted).

ANNEX 2. References

- Bollacker, K., Lawrence, S., and Giles, C. L. 2000. Discovering Relevant Scientific Literature on the Web. *IEEE Intelligent Systems*, 15(2):42-47.
- Bontcheva, K., Dimitrov, M., Maynard, D., Tablan, V., and Cunningham, H. 2002. Shallow Methods for Named Entity Coreference Resolution, *Chaines de references et solveurs d'anaphores*, workshop TALN 2002, Nancy, France.
- Cunningham, H., Maynard, D., Bontcheva, K., and Tablan V. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications, *Proceedings of ACL-02*, Philadelphia, PA.
- De Sitter, A. and Daelemans, W. 2003. Information Extraction via Double Classification, *Proceedings of the ECML/PKDD 2003 Workshop on Adaptive Text Extraction and Mining*, Cavtat-Dubrovnik, Croatia.
- Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A., Shaked, T., Soderland, S., Weld, D., and Yates, A. 2004. Web-Scale Information Extraction in KnowItAll. *Proceedings of the Thirteenth International World Wide Web Conference WWW-2004*.
- Ide, N., Suderman, K. 2004. The American National Corpus First Release. *Proceedings of the Fourth Language Resources and Evaluation Conference (LREC)*, Lisbon, 1681-84.
- Justeson, J. S. and S. L. Katz. 1996. Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text. *Natural Language Engineering*, 3(2), 259-289.
- Kleinberg, J. 1999. Authoritative Sources in a Hyperlinked Environment. *Journal of ACM*, 46(5). p.604-632.
- McCallum, A. 1996. *Bow: A toolkit for statistical language modeling, text retrieval, classification and clustering*, available at <http://www.cs.cmu.edu/~mccallum/bow>".
- Mitkov, R. (forthcoming). The role of corpora in anaphora resolution. In A. Lüdeling, M. Kyto and T. McEney (Eds) *Handbook of Corpus Linguistics*. Mouton de Gruyter.
- Narin, F, Hamilton, K. S., and Olivastro, D. 1997. The Increasing Linkage Between US Technology and Public Science. *Research Policy*, 26(3):317-330.
- Orasan, C. 2003. PALinkA: A Highly Customisable Tool for Discourse Annotation. *Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue*, ACL-03.
- Shinzato, K. and Torisawa, K. 2004. Extracting Hyponyms of Prespecified Hypernyms from Itemizations and Headings in Web Documents. *Proceedings of COLING'04*. Geneva, Switzerland.
- Swanson, D. R.. and Smalheiser, N. R. 1997. An interactive system for finding complementary literatures: a stimulus to scientific discovery. *Artificial Intelligence*, 91:183-203

ANNEX 3. A fragment of the BiRD corpus

```
<corpus>
  <document id="0">
    <document_info>
      <source_url>http://nlp.cs.nyu.edu/app/</source_url>
      <date>18.12.2006</date>
      <permanent_url>
        http://web.archive.org/web/20060423105853/http://nlp.cs.nyu.edu/app/
      </permanent_url>
      <annotation>automatic</annotation>
      <category>NLP resources</category>
    </document_info>
    <template>
      <AREA>parsing</AREA>
      <AREA>grammar</AREA>
      <AREA>NLP</AREA>
      <AREA>terminology</AREA>
      <CREATOR>Keio University</CREATOR>
      <CREATOR>New York University</CREATOR>
      <CREATOR>Satoshi Sekine</CREATOR>
      <CREATOR>Ralph Grishman</CREATOR>
      <CREATOR>Shinichi Torihara</CREATOR>
      <LICENCE>Available by ftp</LICENCE>
      <NAME>Apple Pie Parser</NAME>
      <PLATFORM>Windows</PLATFORM>
      <TARGETLANGUAGE>English</TARGETLANGUAGE>
    </template>
    <entities>
      <entity local_id="1" global_id="1471" type="ORGANIZATION">
        <entity_mention id="1-1">
          <position start="56" end="86" />
          <string text="Department of Computer Science" />
        </entity_mention>
      </entity>
      <entity local_id="2" global_id="111" type="ORGANIZATION">
        <entity_mention id="2-1">
          <position start="135" end="154" />
          <string text="New York University" />
        </entity_mention>
      </entity>
      [...]
    </entities>
  </document>
```