ELECTRONIC WORKSHOPS IN COMPUTING

Series edited by Professor C.J. van Rijsbergen

Jonathan Furner, School of Information and Media Studies, and David Harper, School of Computer and Mathematical Studies, The Robert Gordon University, Aberdeen, Scotland. (Eds)

Information Retrieval Research

Proceedings of the 19th Annual BCS-IRSG Colloquium on IR Research,
Aberdeen, Scotland, 8-9 April 1997

Paper:

On the Marriage of Information Retrieval and Information Extraction

A.M. Robertson and R. Gaizauskas

Published in collaboration with the
British Computer Society

On the Marriage of Information Retrieval and Information Extraction

Alexander M. Robertson

Robert Gaizauskas

Department of Computer Science, University of Sheffield,

Western Bank, Sheffield, S10 2TN..

{sandy,robertg}@dcs.sheffield.ac.uk

Abstract

The techniques of information retrieval and information extraction are complementary, but to date there has been little work aimed at integrating the two. We describe how each of these techniques contributes to the process of transferring information from generator to user, summarise the issues which must be addressed if they are to work together, and report the results of some preliminary experiments on coupling them.

1 Introduction

Information retrieval (IR) identifies documents, from a larger collection, which are (hopefully) relevant with respect to some query. Information extraction (IE) is a technique which processes a document, or collection of documents, to identify pre-specified entities or events. The two techniques are therefore complementary, and their use in combination has the potential to create a powerful tool in text processing, allowing, for example, the automated construction of repositories of structured information from large free text collections.

This paper is organised as follows: first we describe the main features of information retrieval (IR) and information extraction (IE); we then summarise the issues involved in using IR as a filter on texts to be input to an IE system, and finally describe some initial experiments employing queries used to select texts from the World-Wide Web (WWW) for input to an IE system in the subject area of management succession events and labour-union negotiations.

2 Information Retrieval and Information Extraction

2.1 Information Retrieval

2.1.1 Background

Information (or document) retrieval (IR) systems deal with the representation, organisation, and accessing of information items, documents or representatives of documents [1]. IR identifies documents which match a query as presented to the system and which may, or may not, contain the desired information. There are two main approaches in IR, *Boolean* and *ranked-output* or *best-match*.

A Boolean query is constructed using the logical operators AND, OR and NOT. It divides the database being searched into two parts, one containing documents which are considered to be relevant with respect to the query, and the other containing the remaining documents. In the first category, some documents will be more relevant than others and some will not be relevant at all. The same situation is mirrored in the non-relevant set. Within each set, however, there is no differentiation among the documents - they are all considered to be equally relevant, or not. The user must potentially inspect each and every document with no *a priori* knowledge as to where in the set the useful documents

lie. Neither is it possible to predict the likely size of the retrieved set, except with considerable experience of particular systems.

Ranked-output systems rank the documents within a database in decreasing likelihood of relevance with respect to the query. They do this by comparing a set of terms extracted from the query with the sets of terms corresponding to each of the documents in the database. They then calculate a measure of similarity between the query and each of the documents using a numerically-based algorithm and then sort the documents by decreasing degree of similarity with the query. The user can then browse down the list just so far as (s)he considers necessary. This approach takes into account the fact that relevance is not an all-or-nothing matter; it depends not only on the query itself, but must allow for the user's previous knowledge and the items already retrieved and inspected in that search.

In either case, the identification of the required information within the document is the second-stage of a two-stage process, and is normally carried out by the simple expedient of the user reading the document.

2.1.2 *Excite*

To select documents from the WWW, we used *Excite*, one of the many search engines now available, and which had been shown to yield superior results in previous tests [5]. *Excite* claims to have the most accurate and comprehensive indexing based on Intelligent Concept Extraction (ICE), which is 'able to find and score documents based on a correlation of their concepts, as well as actual keywords' (the method used is proprietary and confidential). It is based on an analysis of the entire text of each page, and *Excite* provides an automatically generated summary for each of the 50,000,000 pages that it has indexed. Although we could have used any Web search engine, *Excite* produces its own natural language summaries of the documents which it indexes; not all of the engines do this, and we wanted to compare these summaries with the summaries generated by our information extraction system.

2.2 Information Extraction

2.2.1 *Background*

IR systems are complementary to information extraction (IE) systems. The latter analyse unrestricted texts, which have already been selected, in order to identify pre-specified events, entities or relationships [2, 3, 4]. Thus, IE performs the second task in the two-stage process, i.e., it acts as the "user". While IR outputs require post-processing, the input to an IE system requires pre-processing for maximum effectiveness, i.e., while there is no reason in principle why an IE system should not be fed a heterogeneous collection of texts, it would be computationally inefficient to do so. Examples of IE systems include:

1. Health care delivery: these summarise patient records by extracting diagnoses, symptoms, physical findings, test results and therapeutic treatments [3];
2. Scientific/technical literature monitoring: these extract information about four processing technologies; layering, lithography, etching and packaging, from articles about microelectronic chip fabrication [12];
3. Intelligence gathering: these monitor newswire transcripts of terrorist activities to identify the type of terrorist event, perpetrators, victims and damage to buildings or infrastructure, as well as the time and location of the event(s) [10, 11];
4. Corporate mergers and joint ventures: these extract details of the participating companies, associated products and services, and other details such as the amount of investment capital and the names of the partners [12].

The information extracted may then be used to fill templates, for example to populate the fields in a structured database, or to produce a natural language summary of the text, *inter alia*.

Information extraction is a non-trivial task as there are many ways of expressing the same fact [2], and in addition, information may be distributed across several sentences:

1. *BNC Holdings Inc. named Ms G. Torretta as its new chair-person;*

2. *Nicholas Andrews was succeeded by Gina Torretta as chair-person of BNC Holdings Inc.;*
3. *Ms Gina Torretta took the helm at BNC Holdings Inc. She succeeds Nick Andrews.*

2.2.2 VIE

For our experiments, we used VIE (Vanilla IE system), an IE system developed in our Department. VIE is a research prototype which functions within a language engineering research architecture called GATE – General Architecture for Text Engineering also developed at Sheffield. GATE is a software environment that supports researchers who are working in natural language processing and computational linguistics and developers who are producing and delivering language engineering systems [6, 7]. It is based on the TIPSTER architecture [8], an object-oriented data model designed to support a broad range of document processing tasks and promoted as a standard for the information retrieval and extraction tasks within the ARPA-sponsored TIPSTER text programme.

VIE is a GATE-ified version of the LaSIE (Large-Scale Information Extraction) system [9], Sheffield's entry in the ARPA-sponsored Message Understanding Conference 6 (MUC-6) system evaluations. That is, VIE was derived from LaSIE by standardising LaSIE module interfaces so that all modules communicated with each other via the GATE document manager (allowing for easy substitution of improved modules with similar functionality – e.g., better part-of-speech taggers, or parsers).

The high-level tasks which VIE performs include the four MUC-6 tasks (carried out on *Wall Street Journal* articles):

1. Named entity recognition, the recognition and classification of definite entities such as names, dates, places;
2. Coreference resolution, the identification of identity relations between entities (including anaphoric references to them);
3. Template element construction, a fixed-format, database-like enumeration of organisations and persons;
4. Scenario template construction, the detection of specific relations holding between template elements relevant to a particular information need (in this case personnel joining and leaving companies) and construction of a fixed-format structure recording the entities and details of the relation.

In addition, the system can generate a brief natural language summary of the scenario it has detected in the text.

All of these tasks are carried out by building a single rich representation of the text – the *discourse model* – from which the various results are read off. The system is a pipelined architecture which processes a text sentence-at-a-time and consists of three principal processing stages: lexical preprocessing, parsing plus semantic interpretation, and discourse interpretation. The overall contributions of these stages may be briefly described as follows:

1. Lexical preprocessing reads and tokenises the raw input text, tags the tokens with parts-of-speech, performs morphological analysis, performs phrasal matching against lists of proper names, and builds lexical and phrasal chart edges in a feature-based formalism for hand-over to the parser;
2. Parsing 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', passes on a predicate-argument representation of the current sentence;
3. Discourse interpretation adds the information in its input predicate-argument representation to a hierarchically structured semantic net which encodes the system's world model, adds additional information presupposed by the input to the world model, performs coreference resolution between new instances added and others already in the world model, and adds information consequent upon the addition of the input to the world model.

Further information about VIE or GATE may be obtained by emailing gate@dcs.sheffield.ac.uk, or by visiting our website at

<http://www.dcs.shef.ac.uk/research/groups/nlp/gate>.

2.3 Evaluation

In both IR and IE, effectiveness and efficiency are principally assessed by the twin measures of recall and precision. If a search has retrieved A relevant documents (or filled A template slots correctly) out of the B relevant documents (or B correct slots) in the database, and C documents have been retrieved (or C slots have been filled) in total, then recall and precision are defined to be

$$\frac{A}{B} \times 100 \text{ and } \frac{A}{C} \times 100$$

respectively.

As we have only analysed the results at the level of gross template production, we have adapted the definitions of recall and precision slightly. A is the number of templates generated from relevant documents, and C is the total number of templates generated.

3 IR as Input to IE

There is a clear role for the integration of IR and IE. Indeed, the Tipster research initiative was intended to combine IR and IE, which they refer to as *detection* and *extraction*, respectively [8], but to date the research has largely been carried out independently.

The current level of effectiveness of IR systems means that a proportion of the retrieved documents will not be relevant to the query. IE is relatively computationally-intensive, and it is necessary to ensure as far as is possible that documents used as input are within the topic of the specific IE system. This implies that the query must be biased towards precision rather than recall, and we focus on precision in the experiments reported below. Our investigations have revealed three types of source from which subsets of documents might be drawn:

1. Homogeneous collections from which we may want to draw a subset, i.e., a pre-existing collection of management succession events;
2. Rather more disparate collection in a broader field, e.g., management, but which is likely to contain a reasonable number of relevant documents;
3. Heterogeneous collections, of which the WWW is an extreme example.

The first of these presents a comparatively trivial task, while the second is typical of the type of task which IR is usually asked to perform. The bulk of the experimental work in IR prior to the Text Retrieval Conferences (TREC) has used test collections belonging to particular domains, e.g., LISA - Library and Information Science Abstracts; CACM - computing and INSPEC - electrical and electronic engineering. The third type of source is now of great importance. While IR users tolerate an initial retrieved set which may not match their requirements (and are provided with mechanisms such as relevance feedback in order to improve retrieval) an IE system is only used to maximum advantage if presented with suitable documents; here we want to automate the process as far as is possible.

4 Experimental Procedure and Results

4.1 Background

In order to assess the implications of performing searches on an unrestricted domain for the purpose of identifying documents suitable for input to an IE system, we submitted a query to the WWW. The domain was that of management succession events, as in the example document shown in Section 6. The scenario aims to track changes in company management, and to identify the management post, the company, the current manager, and the reason why the post is or will be vacant, where the new manager came from and where the old manager is going [13]. A relevant article refers to assuming or vacating a post in a company, and must minimally identify the post and either the person assuming the post or the person vacating the post. We want to retrieve documents with information such that templates can be filled with the information. The query reads:

chief executive officer head president chairman post succeed name

This query had previously been constructed to retrieve documents for use in the Sixth Message Understanding Conference (MUC-6) [14]. This MUC queries were intentionally *not* fine-tuned for accuracy, as the intention was to retrieve non-relevant as well as relevant documents. The database which was searched consisted of several million words of newswire articles taken from the Wall Street Journal. The searches were carried out using the IR package *mg* [14, 15]. Thus, the experiments reported here used an initial database of much greater size, *viz*, the 50,000,000 plus documents on the WWW as indexed by Excite.

We used a ranked-output strategy as we anticipated that Excite would retrieve a very large number of documents from the WWW, and we could then process the top n according to the available time. As we have noted, a Boolean query would retrieve an unordered set. The documents that we retrieved were the summaries provided by Excite, and the original documents from which Excite generated these summaries. An example of a document and the templates resulting from processing by VIE are shown in Section 6.

We then used VIE to process the top-50 documents (both the Excite summaries, and the original, indexed documents).

4.2 Results and Discussion

Excite identified no less than 8,847,722 documents as relevant. This is hardly surprising, as an inspection of the query shows that some of the terms such as *name* and *post* are very general ones. We first processed the Excite summaries, concentrating on the top-50 of the summaries, and were encouraged by the overall precision of 68%, as determined by manual inspection. We also calculated the precisions taking account only of summaries in the ranges 1-10, 11-20, 21-30, 31-40 and 41-50; these were 90%, 60%, 70%, 60% and 60%, respectively; thus about two-thirds of the summaries are still relevant in the lower ranks.

VIE produced scenario templates for 12 of the 34 relevant summaries and for none of the 16 non-relevant summaries, i.e recall 34%, precision 100% at the gross level of template production. Fuller analysis of the slot values in the templates has not been carried out as yet. These figures are in line with the LaSIE MUC-6 performance of 37% recall and 73% precision over all slots.

Together with the summaries, Excite provides links to the original documents. As the Web is a dynamically-changing repository, we anticipated that there would be a number of addresses which were no longer valid. In fact no fewer than 21 (42%) of the addresses were no longer valid; in addition, there was one duplicate. Thus only 28 (56%) of the summaries returned by the Excite search were unique candidates for processing by VIE.

VIE generated templates for 21 of the 23 relevant full documents, and three of the five non-relevant ones, yielding recall and precision figures of 91% and 88%.

It was clear why four of the five irrelevant documents had been retrieved, but we were mystified by the fifth; none of the search terms appeared at all; we can only assume that the URL had been reused for a different document in the interval between the Excite indexing and our search.

However, this was a retrospective search in a dynamic environment, and an inspection of the retrieved set shows that the retrieved documents for our query were ephemeral - newswires, press releases, and newspaper articles. We did not place any time restrictions on the search, and the documents were up to two years old. We therefore conclude that the most productive use of an automatic system combining information retrieval and information extraction will be in the trawling of the Web for very recent additions, possibly on a daily basis.

This also emphasises the problems of using Web search engines, especially in some automatic manner. Excite has indexed 50,000,000 documents, but it is clear that not all of the 50,000,000 extant documents are still available (as opposed to the summaries); This problem is not exclusive to Excite, but it does suggest that the designers of search engines will need to devote substantially more resources to the continuous validation of the links on their pages.

5 Conclusions

We have shown that relevant documents for the management succession events can be retrieved from the WWW with a high degree of precision (at least in the top-50 documents) using Excite, and that the templates can be successfully filled by our IE system.

We have also shown that high precision, if somewhat low recall, results can be obtained by running an IE system directly on the output of a Web search. Thus we have demonstrated, albeit somewhat crudely, the principle of coupling IR and IE to derive structured information bases from text on the Web. We also believe that this technique can be refined in order to create an automated system for the comparison of the relative effectiveness of different Web search engines.

There are a number of other issues that must be addressed in creating an integrated IR/IE system:

1. At what point in the ranking does the precision drop so low that further processing by the IE system is unprofitable? Clearly, we cannot process several million documents even if they are all of abstract length;
2. One of the outputs of an IE system is a natural-language summary of the document. How does this compare with the summary provided by Excite, if we process the original document?
3. Is it cost-effective to implement some method of detecting duplicate (or near-duplicate) documents? Ideally, this detection would take place before the documents were processed by the IE system, but this is likely to prove impractical if the size of the database being searched and/or the number of templates were of non-trivial size. We are therefore considering extending the work of Lawson, Kemp, Lynch and Chowdhury [16] to the detection of duplicate documents.

6 Document with Associated Template and Summary

Some of these slots are filled with one of a number of pre-defined values, e.g., VACANCY_REASON must be one of: DEPART.WORKFORCE; REASSIGNMENT; NEW_POST_CREATED or OTH.UNK. Other slots, such as the person's name and the title of the post, are open-ended; these are filled with strings from the text.

Sample Excite Summary

97% Heir Apparent Mandl Leaves AT&T

URL:

http://detroit.thesource.net:80/files/librarywire/96wireheadlines/08_96/DN96_08_19/DN96_08_19_fk.html

Summary: Alex J. Mandl, will leave the largest long-distance telephone company to join a small but ambitious wireless communications firm. Mandl, 52, will become chairman and chief executive officer of Associated Communications, a new unit of The Associated Group, a Pittsburgh-based company with investments in several Mexican wireless companies.

Template

```
<TEMPLATE-ms23-1> :=
    DOC_NR:          "ms23 "
    CONTENT:          <SUCCESSION_EVENT-ms23-34>
<SUCCESSION_EVENT-ms23-35>
<SUCCESSION_EVENT-ms23-34> :=
    SUCCESSION_ORG: <ORGANIZATION-ms23-44>
    POST:           "chief executive officer"
```

```

IN_AND_OUT:                <IN_AND_OUT-ms23-22>
VACANCY_REASON: OTH_UNK
<IN_AND_OUT-ms23-22> :=
  IO_PERSON:                <PERSON-ms23-25>
  NEW_STATUS:               IN
  ON_THE_JOB:               NO
<SUCCESSION_EVENT-ms23-35> :=
  SUCCESSION_ORG: <ORGANIZATION-ms23-44>
  POST:                   "chairman"
  IN_AND_OUT:             <IN_AND_OUT-ms23-24>
  VACANCY_REASON: OTH_UNK
<IN_AND_OUT-ms23-24> :=
  IO_PERSON:                <PERSON-ms23-25>
  NEW_STATUS:               IN
  ON_THE_JOB:               NO
<ORGANIZATION-ms23-44> :=
  ORG_NAME:                 "Associated Communications"
  ORG_TYPE:                 COMPANY
  ORG_LOCALE:               Pittsburgh CITY
  ORG_COUNTRY:              United States
<PERSON-ms23-25> :=
  PER_NAME:                 "Alex J. Mandl"
  PER_ALIAS:                "Mandl"

```

Summary

Alex J. Mandl will become chief executive officer and chairman of Associated Communications.

7 Acknowledgements

We thank Beth Sundheim for supplying the query as used in MUC-6, and Rob Collier, Hamish Cunningham and Kevin Humphreys for comments on an earlier draft of this paper.

References

- [1] Salton, G. and McGill, M.J. Introduction to modern information retrieval. McGraw-Hill, New York, 1983
- [2] Crawford, M. Information extraction [<http://www.dcs.shef.ac.uk/research/groups/nlp/extraction>], 1996. Site visited at 6/1/97
- [3] Lehnert, W. Information extraction. [<http://www-nlp.cs.umass.edu/nlpgroup/nlpie.html>], 1996. Site visited at 13/1/97
- [4] Cowie, J. and Lehnert, W. Information extraction. Commun. ACM 1996; 39:80-91
- [5] Scarlett, J. Implications of the Internet for corporate legal information professionals. MSc dissertation, Department of Information Studies, University of Sheffield, 1996
- [6] Cunningham, H. Gaizauskas, R. and Wilks, Y. A general architecture for text engineering (GATE). CS - 95 - 21 Department of Computer Science, University of Sheffield, 1995. Also available as <http://xxx.lanl.gov/ps/cmp-lg/9601009>

- [7] Gaizauskas, R. Cunningham, H. Wilks, Y. Rodgers, P. and Humphreys, K. GATE – an environment to support research and development in natural language engineering. Proceedings of the 8th IEEE International Conference on Tools with Artificial Intelligence (ICTAI-96), Toulouse, France, pp 58-66, 1996
- [8] Grishman, R. TIPSTER Architecture design document version 2.2, [<http://www.tipster.org>] 1996. Site visited at 7/1/97
- [9] Gaizauskas, R., Wakao, T., Humphreys, K., Cunningham, H. and Wilks, Y. University of Sheffield: Description of the LaSIE system as used for MUC-6. In [13], pp 207-220
- [10] Defence Advanced Research Projects Agency. Proceedings of the Third Message Understanding Conference (MUC-3). Morgan Kaufmann, San Francisco, 1991
- [11] Defence Advanced Research Projects Agency. Proceedings of the Fourth Message Understanding Conference (MUC-4). Morgan Kaufmann, San Francisco, 1992
- [12] Defence Advanced Research Projects Agency. Proceedings of the Fifth Message Understanding Conference (MUC-5). Morgan Kaufmann, San Francisco, 1993
- [13] Defence Advanced Research Projects Agency. Information extraction task: scenario on management succession. In: Proceedings of the Sixth Message Understanding Conference (MUC-6). Morgan Kaufmann, San Francisco, 1995. pp 363-374
- [14] Sundheim, B. Personal communication, 1997
- [15] Witten, I.H., Moffat, A. and Bell, T.C. Managing gigabytes - compressing and indexing documents and images. Van Nostrand Reinhold, New York, 1994
- [16] Lawson, M. Kemp, N. Lynch, M.F. and Chowdhury, G.G. Automatic extraction of citations from the text of English-language patents - an example of template mining. Journal of Information Science 1996; 22:423-436