In *Proceedings of RIAO 97: Computer-Assisted Information Searching on the Internet*, Montreal, Canada, 1997, pp. 356–370.

Coupling Information Retrieval and Information Extraction: A New Text Technology for Gathering Information from the Web

Robert Gaizauskas, Alexander M. Robertson

Department of Computer Science, University of Sheffield, Western Bank, Sheffield S10 2TN, U.K.

Email:{robertg,sandyr}@dcs.sheffield.ac.uk Tel: +44 (0)114 22 21827

Abstract

The techniques of information retrieval and information extraction are complementary, but to date there has been little concrete 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 which indicate that these technologies can be jointly used to construct a structured data resource from free text on the WWW.

Keywords

Document detection; Information retrieval; Information extraction; Text filtering; World-Wide Web

---

1 Introduction

Information retrieval (IR) techniques identify documents from a larger collection which are (hopefully) relevant with respect to some query. *Information extraction* (IE) techniques (also known as *message understanding*) process a document to identify pre-specified entities and the relationships between them and then fill in a structured record or "template" with the identified information. Put another way, IR retrieves relevant documents from collections, IE extracts relevant information from documents. 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. Traditionally IR techniques have been used to search relatively homogeneous document collections (*e.g.* newswires). Even given such homogeneity, IR techniques fall far short of the goal of 100% recall combined with 100% precision [LOSE94, ROBE96]. However, the World Wide Web (WWW) now offers an unprecedentedly large and heterogeneous set of documents for search and retrieval. Many WWW search engines are available, based on various IR techniques, and while they are extremely useful for finding very specific information, they are very limited for general information gathering: it is, for example, not uncommon for queries to produce millions of "relevant" documents. Clearly additional tools are needed to post-process these documents automatically, in order to determine their content and relevance more precisely.

The Natural Language Processing group at the University of Sheffield has been working on IE technology for several years and now, together with its partners in the ECRAN project[1], is applying this technology to document collections constructed by using WWW search engines. This paper describes an experimental setup for constructing WWW-derived document collections using the Excite search engine and for feeding them to the Sheffield VIE information extraction engine in order to produce a structured database of templates. Results of preliminary investigations using this setup are reported. These results:

- demonstrate proof of concept – a structured data resource can be constructed from free text on the WWW;
- suggest that combining IR and IE may provide a useful way for improving query precision;
- have exciting prospects for multilingual document processing, since templates derived by the IE engine can be readily used to generate summaries automatically in different languages (we are currently experimenting in English, French, and Italian and will, in conjunction with a sister EC project, AVENTINUS, be investigating German, Spanish and possible Greek and Swedish).

This paper is organised as follows: first we describe the main features of IR and IE; we then summarise the issues involved in using IR as a filter on texts to be input to an IE system; finally we 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.

[1]ECRAN is a European Community Language Engineering project. Project partners are Thomson-CSF, France; University of Fribourg, Switzerland; Smart Information Services GmbH, Berlin, Germany; University of Roma Tor Vergata, Italy; University of Ancona, Italy; NCSR Demokritos, Athens, Greece; and University of Sheffield, UK

2 Information Retrieval and Information Extraction

2.1 Information Retrieval

2.1.1 Background

Information (or document) retrieval systems deal with the representation, organisation, and accessing of information items, documents or representatives of documents [SALT83]. 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 from atomic query terms (words or phrases) 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, the user will consider some documents to be more relevant than others and some not to be relevant at all. The same situation is mirrored in the non-relevant set. Within each set, however, the IR system makes 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 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, though some systems highlight the query terms in the retrieved text.

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 [SCAR96]. 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 algorithm 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.

2.2 Information Extraction

2.2.1 Background

IE systems are complementary to IR systems. They analyse pre-selected but unrestricted texts in order to identify pre-specified entities, events, and relationships [COWI96, LEHN96]. Thus, IE performs the second task in the two-stage process referred to above, *i.e.*, it acts as does the "user" of an IR system in identifying the required information *within* a retrieved document.

Application domains of IE systems include:

1. Health care delivery: these summarise patient records by extracting diagnoses, symptoms, physical findings, test results and therapeutic treatments [LEHN96];

2. Scientific/technical literature monitoring: these extract information about four processing technologies; layering, lithography, etching and packaging, from articles about microelectronic chip fabrication [DARP93];

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) [DARP91, DARP92] or help international police forces collate information necessary for drug enforcement [AVEN97];

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 [DARP93];

5. Traffic information gathering: these extract details of traffic incidents (time, location, severity of disruption) from police incident logs and formulate advisory broadcasts for motorists [EVAN95];

6. Employment opportunities: these extract details of electronically published job advertisements and construct a job opportunities database accessible by job seekers *via* the Internet in their own language [TREE97].

Typically, an IE system extracts information into a *template*, a user-defined structure which specifies the information to be extracted. An example of a short document, a filled template, and a natural language summary produced from this document may be found in Figure 1. The template in this example was designed to capture information pertaining to management succession events within business organisations; it was defined as part of the MUC-6 final evaluation (see [DARP95], p. 361). Such a template consists of a number of structured objects (*e.g.* `SUCCESSION_EVENT`, `ORGANISATION`) each of which has associated with it some number of slots which must be instantiated by an IE system as it processes a text. Slot fills may be pointers to other objects (*e.g.* `SUCCESSION_ORG`), strings from the text (*e.g.* `ORG_NAME`) or 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`).

Information extraction is a non-trivial task as there are many ways of expressing the same fact [CRAW96], and in addition, information may be distributed across several sentences. For example, each of the following conveys a fact which might be more canonically expressed as *Gina Torretta succeeds Nicholas Andrews as chairperson of BNC Holdings Inc.*

1. *BNC Holdings Inc. named Ms G. Torretta as its new chair-person after Nick Andrews resigned for personal reasons*;

Example document – an 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.
```

A Template filled by VIE

```
<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"
```

A VIE-Generated Summary

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

Figure 1: IE: Inputs and Outputs

2. *Nicholas Andrews was succeeded by Gina Torretta as chair-person of BNC Holdings*;

3. *Ms Gina Torretta took the helm at BNC Holdings Inc. She succeeds Mr Nick Andrews*.

Once a set of filled templates has been obtained from an IE system, it may be put to a variety of uses, for example to populate the fields in a structured database, or to produce a natural language summary of the text, *inter alia*. We discuss applications for template collections further in Section 3 below.

It should be stressed that as yet no general purpose, *i.e.* fully domain independent, IE systems exist. All existing systems require customisation to varying degrees and by persons of varying levels of expertise in order to adapt them to produce results for a new template. The drive to "user-defined IE" [MORG96] is a central activity in current IE research efforts.

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 [CUNN95, GAIZ96]. It is based on the TIPSTER architecture [GRIS97], 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 DARPA-sponsored TIPSTER text programme.

VIE is a "GATE-ified" version of the LaSIE (Large-Scale Information Extraction) system [GAIZ95], Sheffield's entry in the sixth DARPA-sponsored Message Understanding Conference (MUC-6) system evaluation. 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 organisations, persons, places and dates, *e.g.*, *Alex. J. Mandl*, *Associated Communications*, *Gina Torretta*, *BNC Holdings*;

2. Coreference resolution, the identification of identity relations between entities (including anaphoric references to them), *e.g., Ms Gina Torretta* and *she* in Number 3 above;

3. Template element construction, the completion of a fixed-format, record-like structure for all organisations and persons, *e.g.,* the `ORGANIZATION` and `PERSON` structures in the template in Figure 1;

4. Scenario template construction, the detection of specific relations holding between template elements relevant to a particular information need and construction of a fixed-format structure recording the entities and details of the relation, *e.g.,* the overall template in Figure 1.

In addition, the system can generate a brief natural language summary of the scenario it has detected in the text, as shown at the foot of Figure 1.

VIE carries out all of these tasks by using NLP techniques to build 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 one 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 their 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.

Aspects of the system's processing are described in more detail in [GAIZ95] and [GAIZ97][2].

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* corrects 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

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

respectively.

Both IR and IE techniques have been the subject of extensive evaluation efforts through the DARPA Text Retrieval Conferences and Message Understanding Conferences (see, *e.g.* [TREC95] and [DARP95]). While the experimental setup of these conferences is quite elaborate the evaluation measures revolve centrally around these twin measures.

3 IR as Input to IE

There are clearly applications of IE technology in monitoring newswires, analysing large text bases such as patient records, academic journals, or police reports, and in information gathering from

[2] VIE and GATE are both freely available for research purposes. Email `gate@dcs.sheffield.ac.uk`, or visit our website at `http://www.dcs.shef.ac.uk/research/groups/nlp/gate` to find out how to obtain them.

the WWW. However, IE techniques are computationally intensive[3] and therefore it is necessary to ensure as far as is possible that documents input to an IE system are likely to fall within the domain of the specific IE system.

Thus, it is natural to think of ways in which an IR system can be used as a front-end to an IE system to retrieve, from the source collection, documents which are relevant to an extraction scenario. Coupling IR and IE in this way is not a novel idea. Indeed, the TIPSTER research initiative was intended to combine IR and IE, to which it refers as *detection* and *extraction*, respectively [GRIS97], but to date the research has largely been carried out independently and to our knowledge few concrete studies have been made that involve coupling real systems.

At the conceptual level the linkage of these systems is as pictured in Figure 2. A user submits a query *via* an IR system to the chosen document database. The retrieved documents form a subset of this original document set and are then passed to the IE system which extracts information from them in the form prescribed by the user *via* a template definition. The resultant templates can then be utilised by any number of further application programs.

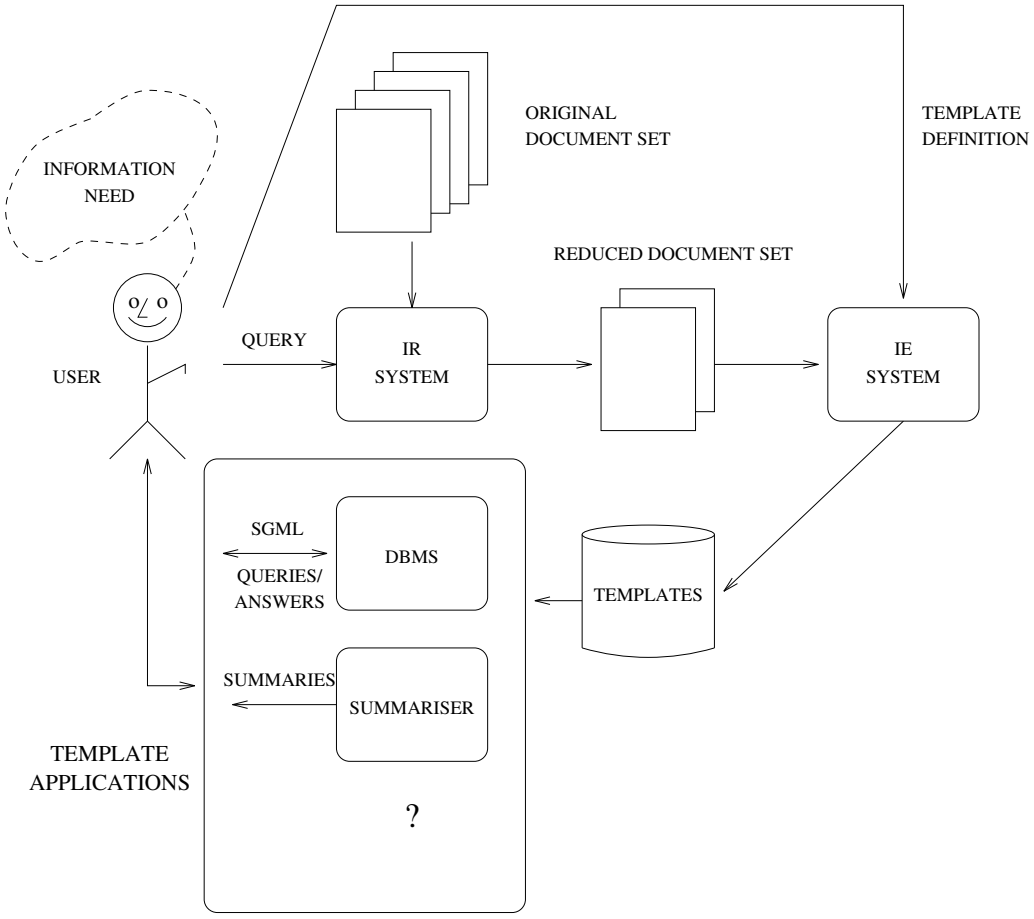

Figure 2: Conceptual architecture of a coupled IR-IE system

The obvious uses to which the templates may be put are database querying and natural language

[3]IE systems at MUC-6 processed text roughly in the range of 0.1 - 45 kilobytes/minute. While such systems might just about keep up with the input of a single newswire, they are clearly, at present, incapable of dealing with extremely large and dynamic text sources like the WWW.

summarisation (possibly including translation). However, there are further possible applications. Data mining could be carried out against the template collection (especially if the template database is maintained over some historical period, in which case previously undiscovered correlations might be revealed, *e.g.*, between factors affecting commodity price fluctuations). Or, the templates could be used to enhance the performance of the IR system. For example, the reduced document set which has been processed by the IE system could be further filtered by discarding all documents in it which have failed to produce sufficiently instantiated templates. This use of IE systems to perform *text-filtering* is discussed by Lewis and Tong [LEWI92] and has become a regular feature in the MUC evaluations (some irrelevant texts are deliberately supplied to the IE systems to see how well the systems can detect them[4]). Of course this procedure would slow retrieval considerably, but for some applications this could be tolerable if the resulting gains in precision were substantial.

IE can also be used to improve IR by providing possibilities for more refined indexing. For example, indexing pre-categorised names could be useful (so, *e.g.*, occurrences of *Ford* as a company are indexed distinctly from those of *Ford* as a person and those of *Ford* as a place). So too could indexing terms within relations – for example, it could be useful to be able to index companies that stood in the role of predators in corporate acquisition events (see [PIET97] for an example of such an application of IE to IR in the legal domain).

4 Experimental Procedure and Results

4.1 Background

In order to assess the merits of coupling an IE system to an IR search engine with access to the WWW, we employed the same query as was used to gather documents for the MUC-6 system evaluation and the same template that was used to define the extraction task. The domain was management succession events, for which articles similar to the example document shown in Figure 1 are relevant. The IE scenario template was designed 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 [DARP95]. 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. The query reads:

chief executive officer head president chairman post succeed name

As noted, this query was constructed to retrieve documents for use in the Sixth Message Understanding Conference (MUC-6) [SUND96] and was deliberately intended to return a considerable number of irrelevant documents, since the participants were asked to undertake text filtering as well as information extraction. The database to be 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* [SUND96, WITT94]. The MUC-6 training and test sets consisted of 100 documents of which roughly half were relevant and half were non-relevant.

However, 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.

[4]IE systems in MUC-6 can perform text-filtering to a high degree of accuracy, with the recall being as high as 98% and precision being as high as 96%, though the usual inverse relationship between these measures appears to hold. However, the proportion of relevant documents is much higher in MUC than is the case in a typical IR task and very much higher than is the case in a set of documents retrieved from the WWW.

4.2 Procedure

We adopted the following procedure. First we submitted the above query to Excite. Since Excite produces ranked-output we were then able process the top *n* retrieved documents according to the available time. In this experiment we chose to examine the top 100 hits only. In addition to links to the original documents, Excite produces short summaries of the documents which are returned as part of the search results. These were removed from the search results and saved as text files. References to the full documents were identified in the search results and the full documents were also retrieved and stored as text files. GATE document collections for both the summaries and the full documents were built and then VIE processed the two document collections. VIE produces null templates for documents which it determines do not meet the MUC-6 minimal content requirements and it produces a completed template and a natural language summary for documents which it does deem relevant (refer to Figure 1 for an example of an Excite summary and the template and summary resulting from processing it by VIE).

Both the summaries returned by Excite, and the full documents subsequently retrieved, were manually examined and judged for relevance by applying the MUC-6 criteria for minimal required content. This is, in one sense, an unfair criterion to apply to the Excite summaries, since for a document which reports, but is not predominantly concerned with, management succession events, a summary might quite reasonably omit them (though it is not clear to what extent Excite summaries are intended to be geared to the domain; *i.e.*, it is not clear whether for a given document Excite generates the same summary regardless of the domain, or whether the summary is tailored in some way to the domain). However, this analysis is of interest to us, as it gives an indication of the extent to which an IE system might get by with processing summaries produced by a search engine.

Precision figures for relevancy of the Excite summaries and documents were obtained straightforwardly from the above manual analysis. Precision and recall figures for the VIE templates for summaries and full documents were then determined by treating the Excite-retrieved documents as a closed set.

4.3 Results

The query stated above was given to Excite on 04/02/97 at 18:20 GMT and resulted in 2,868,989 hits. 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 examined the top 100 hits. Eight of these could not be used - three were duplicates and five of the URLs were no longer valid when we attempted to download the full documents. Thus our test set consisted of 92 summaries and 92 documents.

Manual analysis indicated that 72 (78%) of the full documents were relevant (*i.e.* contained the minimal information to warrant the production of a management succession event template as per MUC-6 rules). We also divided the ranked documents into ten groups consisting of the highest ranked tenth, next highest ranked tenth, and so on, and calculated the precisions in each of these tenths. The rationale for this procedure was to determine how fast precision (with respect to the template filling task) was dropping off even within the top 100 documents. The precision figures for these ten groups are: 90%, 80%, 80%, 70%, 40%, 70%, 90%, 90%, 60%, and 60%, respectively. Thus some two-thirds of the documents are still relevant even at the lower rankings. Given the large number of documents retrieved by Excite, this finding is perhaps not too surprising. It also shows that further work is required to determine an appropriate cut-off point in the relevancy ranking for passing documents onto the IE system.

Our manual analysis also revealed that 54 (59%) of the summaries were relevant, by the MUC-

6 criteria. The corresponding precisions for each tenth of the ranking are: 50%, 60%, 70%, 40%, 60%, 60%, 70%, 80%, 20% and 30%, respectively. Not only are the percentages lower, but the summaries themselves are less appropriate for this task, as we demonstrate below.

VIE generated templates for 67 of the 72 relevant full documents and 11 of the 20 non-relevant ones, yielding recall and precision figures of 93% and 86%. Against the summary collection, VIE generated templates for 21 of the 54 relevant documents, and 3 of the 38 non-relevant ones, yielding recall and precision figures of 39% and 88%, respectively.

The total sizes (in words [5]) of the respective document collections are given in Table 1.

Document Type	Number	Size	Avg. Doc. Size
Full documents	92	51014	555
Excite summaries	92	5694	62
VIE summaries generated from the full documents	78	3173	41
VIE summaries generated from the Excite summaries	24	340	14

Table 1: Sizes of the various document collections

No attempt was made to assess the accuracy of the information in the VIE-extracted templates. Quantitative measures of the precision and recall for template objects and slots were made for LaSIE (the system from which VIE was derived) on blind test data at MUC-6. The official system scores there were 37% recall and 73% precision (combined recall and precision of 49%) and while the system has been modestly enhanced since then, there is no reason to believe that the template filling scores have changed dramatically.

4.4 Discussion

The first point to be made is that the above experiment shows how a structured database – a collection of templates – can be derived automatically from unstructured information on the Web. Due to its slowness[6] an IE system such as VIE could never derive this information by crawling the Web itself; however, coupled with an IR engine to filter the mass of irrelevant text, such processing becomes feasible.

Secondly, it can be seen that an IE system such as VIE could serve as a useful text filter on the output of an IR system, particularly with respect to processing the summaries only. If the full text of the 92 documents identified by the IR system were passed to the IE system before the user viewed the output, then for some loss of retrieved relevant documents (7%) the user gains a marginally greater amount in precision (8%) . The effects are clearer with the summaries: a loss of 43% of retrieved relevant documents results in a 29% increase in precision. An earlier experiment on a smaller set of Excite summaries reported in [ROBE97] showed even more dramatic effects, with precision rising by 32% to 100%, though at the cost of about 65% of the retrieved relevant documents. Further experimentation is necessary to determine whether these results are reliably repeatable.

Thirdly, an IE system VIE allows a very different sort of summarisation to that provided by abridgement techniques such as that used by Excite. To see this, let us consider one example in somewhat more detail.

[5] As reported by the Unix utility *wc*.

[6] Without any real attention to optimisation, VIE fully processes about 1 kilobyte of text per minute.

One of the 92 full documents retrieved concerns the appointment by GM of its president to the post of chairman. This text was 447 words long – about 80% of average length – and consisted of 20 sentences (too long to include here). Excite produced a 64 word summary consisting of all of one, and part of another sentence from the original text. We reproduce it here to give the flavour of the approach:

> *Summary: "Now, some three years later, it's clear that GM's management team under Jack Smith's leadership has turned GM around," Smale, a former chairman of Procter & Gamble Co. "The changes announced today will permit (Smale) to continue the leadership role he has played on the GM board, while permitting him to reduce his day-to-day involvement in GM's governance," Smith said in the announcement.*

Note that the first "sentence" is not actually a sentence (the original has been truncated and has lost its main verb. The major event reported in the article – the appointment by GM of Smith to chairman has not been captured.

VIE, by contrast, generates rather more terse sentences as can be seen in its 114 word summary of the same original document.

> *General Motors Corp. appoints Jack Smith as president.*
> *Jack Smith will become Chairman, chairman and executive of General Motors Corp..*
> *General Motors Corp. heads John G. Smale as Chairman, chief executive and executive.*
> *John G. Smale became chairman of General Motors Corp..*
> *Harry J. Pearce will become vice chairman, executive, President and chairman of General Motors Corp..*
> *Harry J. Pearce joins General Motors Corp. as executive vice president.*
> *Harry J. Pearce served as general counsel of General Motors Corp..*
> *John D. Finnegan will succeed Heidi Kunz as chief financial officer of General Motors Acceptance Corp.*
> *Heidi Kunz resigned as chief financial officer of ITT Industries.*
> *Heidi Kunz joins ITT Industries as chief financial officer.*

While VIE captures the major event and a number of ancillary management succession events correctly, some of the other assertions here are wrong and others impossible (Kunz resigning and joining). These are due to defects of the linguistic processing in the system, and work is ongoing to overcome these weaknesses.

The focused approach of VIE can be seen here; it is dedicated to performing one task – extracting facts about management succession. The Excite summary is indicative, *i.e.*, it is clear that something is happening among the management at GM, but without an inspection of the original source, it is not certain what that is. It has been generated by selecting putatively pertinent sentences from the original text, while the VIE-generated summary is more concise, and not generated from the text. However, we emphasise that a direct comparison is perhaps not appropriate, as the two summarisation techniques are aimed at quite different objectives. In addition, although the VIE summary is more focused, it is also the only type of summary that it can generate in its current configuration.

One final feature of the VIE summary should be mentioned. Since it it generated from a fixed format template by relatively crude natural language generation techniques, it is easy to generate summaries from the template in different languages. This raises the possibility of doing information extraction from documents in one language and generating summaries of them in another. In the context of a resource like the Web with its highly multilingual user population, this is a real attraction and we are actively investigating this avenue now [AZZA97].

Finally, it is worth noting that using the same IE system as a postprocessing filter on multiple IR systems might give some indication of the relative merits of these systems. If we assume that the IE system will maintain roughly constant recall and precision rates as a text-filter within its domain, then given the top n documents from each of two IR systems, the IR system for which the IE system proposes the most templates should be producing more relevant documents in the top n. This could provide a technique for automatic assessment of IR systems.

5 Conclusions

We have shown that high precision and good recall can be obtained by running an IE system directly on the output of a WWW search. Thus we have demonstrated the principle of coupling IR and IE to derive structured information bases from text on the WWW. We have also shown that the best results can only be obtained if the original source is processed, rather than the abridgement-type summary generated by a search engine (indeed, VIE failed to generate a summary at all from the Excite summary of the example document discussed in the previous section).

Clearly, what we have shown here concerning the linking of these technologies is initial experimentation only. Amongst the issues that remain to be addressed in creating an integrated IR/IE system are:

1. At what point in the IR system's relevancy 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. 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 [LAWS96] to the detection of duplicate documents.

Improvements are required, and can be expected, independently in both IR and IE; however with their coupling a new era in text technology may well be about to begin.

6 Acknowledgements

We thank Beth Sundheim for supplying the queries as used in MUC-6, and Rob Collier, Hamish Cunningham, Kevin Humphries and Shaaron Ainsworth for comments on drafts of this paper. We also thank the UK EPSRC and the European Commission for funding which has made the development of VIE/LaSIE and GATE possible.

References

[AVEN97] AVENTINUS: advanced information system for multinational drug enforcement. [http://www2.echo.lu/langeng/en/le1/aventinus/aventinus.html], 1997. Site visited at 10/5/97

[AZZA97] Azzam, S., Humphreys, K., Gaizauskas, R., Cunningham, H. and Wilks, Y. Using a language independent domain model for multilingual information extraction. In Proceedings

of the IJCAI-97 Workshop on Multilinguality in the Software Industry: the AI Contribution (MULSAIC-97), 1997.

[COWI96] Cowie, J. and Lehnert, W. Information extraction. Communications of the ACM 1996; 39:80-91

[CRAW96] Crawford, M. Information extraction. [http://www.dcs.shef.ac.uk/research/groups/nlp/extraction], 1996. Site visited at 6/1/97

[CUNN95] 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`

[DARP91] Defence Advanced Research Projects Agency. Proceedings of the Third Message Understanding Conference (MUC-3). Morgan Kaufmann, San Francisco, 1991

[DARP92] Defence Advanced Research Projects Agency. Proceedings of the Fourth Message Understanding Conference (MUC-4). Morgan Kaufmann, San Francisco, 1992

[DARP93] Defence Advanced Research Projects Agency. Proceedings of the Fifth Message Understanding Conference (MUC-5). Morgan Kaufmann, San Francisco, 1993

[DARP95] 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

[ECRA97] ECRAN: Extraction of content: research at near market. http://www2.echo.lu/langeng/en/le1/ecran/ecran.html], 1997. Site visited at 10/5/97

[EVAN95] Evans, R., Gaizauskas, R., Cahill, L., Walker, J., Richardson, J. and Dixon, A. POETIC: a system for gathering and disseminating traffic information. Journal of Natural Language Engineering 1995; 1:4, 363-387

[GAIZ95] 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 [DARP95], pp 207-220

[GAIZ96] 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

[GAIZ97] Gaizauskas, R. and Humphreys, K. Using a semantic network for information extraction. Journal of Natural Language Engineering. In press.

[GRIS97] Grishman, R. TIPSTER architecture design document version 2.2, [http://www.tipster.org] 1996. Site visited at 7/1/97

[LAWS96] 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

[LEHN96] Lehnert, W. Information extraction. [http://www-nlp.cs.umass.edu/nlpgroup/nlpie.html], 1996. Site visited at 13/1/97

[LEWI92] Lewis, D.D. and Tong, R.M. Text filtering in MUC-3 and MUC-4. In: Defence Advanced Research Projects Agency. Proceedings of the Fourth Message Understanding Conference (MUC-4). Morgan Kaufmann, San Francisco, 1992, 51-66

[LOSE94] Losee, R.M. Upper bounds for retrieval performance and their use for generating optimal Boolean queries: can it get any better than this? Information Processing and Management 1994; 30:193-203

[MORG96] Morgan, R.G. An architecture for user defined information extraction. Technical report 8/96, Dept. Computer Science, University of Durham, 1996

[PIET97] Pietrosanti, E. and Graziadio, B. Artificial intelligence and legal text management: tools and techniques for intelligent document processing and retrieval. In: Natural Language Processing: Extracting Information for Business Needs, Unicom Seminars Ltd., Uxbridge, UK, 1997, pp. 277-291

[ROBE96] Robertson, A.M. and Willett, P. An upperbound to the performance of ranked-output searching: optimal weighting of query terms using a genetic algorithm. Journal of Documentation 1996; 52:405-420

[ROBE97] Robertson, A.M. and Gaizauskas, R. On the marriage of information retrieval and information extraction. In: Furner. J. and Harper, D.J. eds. Information retrieval research 1997: proceedings of the 19^{th} annual BCS-IRSG colloquium on IR research, Aberdeen, Scotland. London, Springer-Verlag, 1997. In press.

[SALT83] Salton, G. and McGill, M.J. Introduction to modern information retrieval. McGraw-Hill, New York, 1983

[SCAR96] Scarlett, J. Implications of the Internet for corporate legal information professionals. MSc dissertation, Department of Information Studies, University of Sheffield, 1996

[SUND96] Sundheim, B. Personal communication, 1996

[TREC95] Harman, D.K. ed. Proceedings of the Fourth Text Retrieval Conference (TREC-4). Washington, National Institute of Standards and Technology, NIST Special Publication 500-236, 1995

[TREE97] TREE: Trans European Employment [http://www2.echo.lu/langeng/en/le1/tree/tree.html], 1997. Site visited at 10/5/97

[WITT94] Witten, I.H., Moffat, A. and Bell, T.C. Managing gigabytes - compressing and indexing documents and images. Van Nostrand Reinhold, New York, 1994