Challenges for chemical information extraction and text retrieval

Bas van Bakel*

Enschede, The Netherlands

Geert Postma

Nijmegen, The Netherlands

1. Introduction

Computer and information technology have expanded enormously over the last few years, creating a number of promising possibilities regarding (rapid) exchange of information. New concepts like Gopher, ftp, URL, http, and html are now part of required standard knowledge when dealing with computer-mediated communication (cf. [1,2]). The optimism is widespread that these new technologies already contribute to a more effective and efficient way of conducting scientific research. We will try to show that this view is somewhat too optimistic as we focus on vital issues concerning automatic document processing.

2. Document processing

Until now, new media such as the World Wide Web (WWW) and CD-ROM have been used mainly for conventional information distribution, although in a digitalized form [2]. Electronic distribution of chemical information, for example, is usually provided by existing institutions such as universities, publishing companies, etc. However, the number of documents available through digital media is growing enormously. In science alone, hundreds of thousands of publications within various disciplines are released in computer-readable form every year. Because of this huge supply of information it is extremely difficult – if not impossible – to select from all these documents (of which just a subset actually consists of written articles)

the ones that are of interest, and thus to keep track of the developments in specific scientific disciplines (or all disciplines, which would be the task of an electronic library service). This is what has brought about the demand for computer systems that can track down information of interest and make it available and accessible, and has resulted in various initiatives to develop computer programs that are known as Information Retrieval Systems (or Document Retrieval Systems) and Information Extraction Systems.

The main goal of Information Retrieval Systems (known examples are the internet Search Engines) is to locate and open up relevant information within documents following a special request by a user. Information Extraction Systems are concerned with extracting all relevant information from previously selected documents and storing this information in a database, which can then serve as a knowledge resource for scientists (for example a source of analytical methods and techniques for chemists). Both kinds of systems may contribute to a more effective and efficient way of information acquisition. They may even prove to be a substantial factor in future developments on electronic publishing (cf. Ref. [3]), as the role, and with it the viability, of (electronic) publishers might be reduced considerably once scientists have good information retrieval and extraction systems at their disposal. However, this is an image of a still distant future.

The majority of current information retrieval systems are based mainly on statistical methods. Other knowledge, like knowledge on the domains (e.g. Analytical Chemistry) that are covered by the documents, or knowledge on natural language processing, is only used to a small extent. The scientific field of Natural Language Processing (NLP) still has to make its first significant contribution to improving document retrieval systems (cf. Ref. [4]). Nevertheless, the possibilities of statistics-based programs are limited (cf. Ref. [5]). At TREC (Text Retrieval Conference) an annual conference organised by the American National Institute of Standards and Technology, state-of-the-art

*Corresponding author.
E-mail: bakel@cs.utwente.nl

document retrieval systems have reached an effectiveness of 30% recall (defined to be 100% when all relevant documents to a specific user request are found and 0% when none are found), at a precision of 55% (defined to be 100% when only relevant documents are found, and 0% when no document is relevant). Whether these results are satisfying depends on the objectives pursued. The results may suffice for someone interested in a few introductory documents or links to other documents on a specific topic, but for a patent office, for example, they are hardly sufficient. What is more, regardless of the objectives of retrieval, any search engine that produces 70% of irrelevant documents will not be highly appreciated. Present search engines produce a great amount of digital 'scrap', which is disadvantageous for both the users of these systems (it leaves a lot of work to be done) and their producers. Nobody wants to pay for digital scrap, or systems that produce it.

Automatic extraction of information deals to some extent with similar problems as information retrieval. Both involve text analysis of some kind, and both may benefit from the use of knowledge of

the domain(s) involved, and NLP knowledge. Like in retrieval, information extraction systems based on statistical methods alone perform insufficiently. In extraction, 100% coverage with 100% accuracy cannot be obtained. In some cases it is even puzzling to human experts to fully understand what is meant exactly. The performance of current NLP-based information extraction systems indicate that further research is still necessary. A comparison of the performance of a series of systems in 1991 revealed that for relatively free text (news articles on a specific subject) an average of 26% recall and 52% precision could be obtained [6]. A more recent investigation [7] showed figures of 46 and 52%, respectively. These figures were the means of the three best performing systems on one domain, but cannot directly be compared with the older figures, because the information extraction task that resulted in the latter figures was evaluated to be more difficult, partly because the metrics used changed slightly. On the same corpus four human analysts obtained 77% recall and 79% precision. These figures are not satisfying. Further research is needed to build systems that perform better. Like in Document Retrieval, research should involve development of deterministic rather than probabilistic systems. This is why much effort is put into projects that employ NLP knowledge and domain knowledge. At present, various research projects of an interdisciplinary nature are being conducted, involving not only software engineers but also knowledge engineers, domain experts and NLP scientists. In the long run these projects are expected to result in knowledge-based systems that perform well in real situations (e.g. as Search Engines for the internet), and replace the current ones.

3. Chemical information processing

Within the domain of chemistry, several research projects have been involved in automatic information extraction covering various subdomains. For example, Nishida et al. [8] developed a system for extraction and storage of information contained within patent claim sentences in the domain of semi-conductor production. Ai et al. [9] developed a system for the extraction of (part of the) procedural synthesis information from the experimental section of a journal for organic chemistry. Crowdhury and Lynch [10] have worked on a system for extraction, representation and storage of textual

Fig. 1. Overview of the Nijmegen Information Extraction System.

descriptions of compounds in a chemical database, and the work of Mars and Van der Vet [11,12] is concerned with information extraction from a set of document descriptions on mechanical properties of ceramic materials.

At the University of Nijmegen, the Netherlands, we have developed a system for information extraction from short analytical method descriptions (title plus abstract) taken from *Analytical Abstracts*, on four analytical techniques, i.e. High-performance liquid chromatography, inductively coupled plasma, atomic absorption spectrometry, and titrimetry. These document descriptions contain relatively free text within the analytical chemistry domain. In addition, part of the sentences are written in, more or less, telegram style, and the sentences may contain a number of defined abbreviations. Although abstracts do not seem to be a reliable source of information [13] they were used because of the specific text characteristics being an ideal test domain of the system for various applications. The information that should be extracted are the characteristics of analytical methods (analyte, matrix, working range, applied technique, precision, accuracy and detection limit; see also the content requirements of analytical abstracts as described in Ref. [13]) and the described actions that (roughly) comprise the analytical method

together with the participants and circumstances of the actions.

The information extraction system consists of two major components: a linguistic module and a chemical knowledge module. A strict modular approach is used for reasons of maintainability and expendability. Extraction of relevant information is conducted as follows. First, a domain-independent semantic analysis of the abstract is produced, by identification and analysis of word classes, syntactic structures and thematic relations within all sentences in the text, respectively. After this, analytical chemical discourse analysis is performed to extract all relevant information, by the chemical knowledge module. In Fig. 1 this process is illustrated. Discourse (or pragmatic) analysis consists of constructing the 'story' that is told in the abstracts, using background information of the various analytical techniques that are mentioned in the abstracts. This analysis is based on Sowa's Conceptual Graph Theory (cf. [14]), which seems to be unlimited in representation possibilities of conceptual structures. An example of the system's output is presented in Fig. 2. The linguistic module is based on principles of Chomsky's Government and Binding Theory (cf. [15]). This theory is used in other systems for other domains as well (cf. [16–20]). Fig. 1 illustrates that the system is domain-independent in its set-up, which means that its parts can be used to cover English texts on a different domain, or analytical chemical texts written in a different natural language.

The above is just a very brief description of the information extraction system that was developed. In Refs. [21–23] the system is described in great detail.

4. Discussion

In four years of research we have built a knowledge-based, deterministic system that can conduct automatic information extraction from analytical chemical document descriptions. The linguistic module is capable of performing a thematic analysis of the input, provided that sufficient information on words, word classes and (in case of verbs and nouns) thematic roles is available. In addition, a discourse analysis module together with a knowledge base was developed that can construct the stories told in the document descriptions. However, so far the entire system only works on a small body of texts, although the lexical and syntactic analyser

Fig. 2. Result of the interpretation of the title phrase 'Determination of phosphorus in milk by electrothermal-atomization atomic-absorption spectrometry with L'vov platform and Zeeman background correction.' The extracted parts of information are presented in italics, together with their appropriate type or storage field.

cover a great deal more of the same body. Initially, the goal of the Nijmegen research was to use a body of 124 texts, and to divide it in two subsets, developing the system using one subset, and testing it with the other. This meant that the system should be able to process 62 abstracts. During the research this goal appeared to be too ambitious, most of all because of the limited time and manpower available. We were facing a number of problems that are exemplary for common problems in document processing systems, some of which we will discuss below.

Previous to the research discussed here, two pilot studies were conducted (cf. [24,25]) in which the possibilities of building an information extraction system based on the use of domain and linguistic knowledge were explored. We assumed that if we elaborated on the findings of these pilot studies, it would be possible to build a working implementation for the entire body of texts within four years, i.e. based on an investment of eight man-years only. However, this turned out to be too optimistic. During the research we were confronted with a number of practical problems which we had not anticipated, or had underestimated, which hindered the development of a large-scale operative system. For example, in order to process a large number of document descriptions, the system should have a large lexicon containing all the required syntactic and semantic information in a processable format. We had to build such a lexicon ourselves, as any existing source was either unusable or too expensive. The same also holds for knowledge bases. Scientific publishers, or institutions like Chemical Abstracts Service, could play a role of importance in this respect. These difficulties regarding knowledge resources are a few of the main problems when building reality-level document processing systems. Creating these facilities is not only expensive because it is labour-intensive, but also because the required knowledge resources can only be obtained at considerable costs. Obviously, this should not necessarily be a problem for commercial R&D teams.

Other problems we encountered during the research involved the quality of document descriptions. Misspelled words, ungrammatical sentences and even missing information on the characteristics of the analytical methods described (despite all the criteria that these document descriptions have to meet) complicated automatic information extraction considerably. Given the limited amount of time available, we were forced to deviate from the

original research objectives. We therefore decided to focus on finding out exactly what (kind of) knowledge should be used in the information extraction system, and in what way, in other words to follow a principle-based approach. The system thus built should first and foremost have sound theoretical foundations rather than be based on relatively ad hoc analysing strategies. As a consequence of this change of objective, we had to abandon the original objective of processing 62 document descriptions and focus on a small body of representative descriptions instead. This task proved to be far from simple, as principle-based automatic information extraction proved to be virgin territory. Despite these setbacks, a system was built based on a set of principles that prove to be fundamental for linguistics-based and knowledge-based information extraction. For example, a strategy was developed which does away with the notorious problem of syntactic and semantic ambiguities within texts. The combination of linguistic and domain knowledge has proved to be particularly powerful in this respect.

The system produced can be regarded as a good starting point for building real-world applications of document processing systems. This fact is acknowledged by another research group: the Knowledge-Based Systems Group of the University of Twente uses parts of the linguistic module in Condorcet, a domain-specific Information Retrieval system for the domains of mechanical properties of engineering ceramics as a subfield of engineering, and epilepsy as a subfield of medicine. In this project, the problems sketched here are tackled, as the main objective is to build a prototype reality-level IR system that is robust and fast, and performs better than existing state-of-the-art IR systems.

5. Conclusions

Information technology has still a long way to go before document processing systems are fit for the job we want them to do. Systems based on statistical methods have their limitations, and principle-based commercial applications for real life situations still have to be developed. However, there is no reason to be too pessimistic about this. At a recently held conference on full-text processing in Bath, UK [26], producers of commercial text search engines unanimously acknowledged that the recent past has shown that million dollar investments in manpower and computer systems do not

guarantee well-performing commercial products. What is more, it is now also a generally accepted view that future systems should employ NLP-based and knowledge-based techniques. It is therefore very likely that investments will be made in the development of such systems, and possibly concurrently in projects as discussed above, although the objectives will most likely be more commercially oriented. Perhaps this will lead to good document processing systems that will perform considerably better than we now dare hope. Only then can we truly say that the new technologies have contributed to more effective and efficient scientific research.

References

- [1] S.M. Bachrach, Trends Anal. Chem., 14 (1995) 182.
- [2] B.M. Tissue, Trends Anal. Chem., 14 (1995) 426.
- [3] S.R. Heller, Publishing on the Internet: a Proposal for the Future, 28 July 1995.
- [4] D. Harman, P. Schäubele and A. Smeaton, in R.A. Cole, J. Mariani, H. Uszkoreit, A. Zaenen and V. Zue (Editors), Survey of the State of the Art in Human Language Technology, Center for Spoken Language Understanding, Oregon Graduate Institute.
- [5] K.S. Jones, Information Access, lecture at the Language and Technology Awareness Day, organised by the European Committee, Prague, 18 November 1994.
- [6] W. Lehnert and B. Sundheim, AI Mag., 12 (1991) 81.
- [7] Proc. of the Fifth Message Understanding Conference (MUC-5), Baltimore, MD, August 25–27, 1993, Morgan Kaufmann, San Francisco, CA, 1993.
- [8] F. Nishida, S. Takamatsu and Y. Fujita, J. Chem. Inf. Comput. Sci., 24 (1990) 163.
- [9] C.S. Ai, P.E. Blower and R.H. Ledwith, J. Chem. Inf. Comput. Sci., 30 (1990) 163.
- [10] G.G. Crowdhury and M.F. Lynch, J. Chem. Inf. Comput. Sci., 32 (1992) 468.
- [11] N.J.I. Mars and P.E. van der Vet, in H. Czap and W. Nedobity, Editors, TKE '90: Terminology and Knowledge Engineering, Indeks Verlag, Frankfurt am Main, 1990, pp. 352–362.
- [12] P.E. van der Vet and N.J.I. Mars, J. Chem. Inf. Comput. Sci., 33 (1993) 564.
- [13] G. Postma and G. Kateman, Anal. Chim. Acta, 265 (1992) 133.
- [14] J.F. Sowa, Conceptual Structures: Information Processing in Mind and Machine, Addison-Wesley, Reading, MA, 1984.
- [15] N. Chomsky, Lectures on Government and Binding, Foris, Dordrecht, 1981.
- [16] J. Fargues, M.-C. Landau, A. Dugourd and L. Catatch, IBM J. Res. Dev., 30 (1986) 70.
- [17] M.L. McHale and S.H. Myaeng, Knowl.-Based Syst., 5 (1992) 213.
- [18] M. Schröder, in H.J. Ohlbach, Editor, GWAI-92: Advances in Artificial Intelligence, Lecture Notes in Artificial Intelligence 671, Springer-Verlag, Berlin, 1992, pp. 221–234.
- [19] P. Zweigenbaum, Editor, Consortium Menelas, Menelas: An Access System for Medical Records using Natural Language, Computer Methods and Programs in Biomedicine, Special issue of AI Mag., 45 (1994) 117–120.
- [20] A.-M. Rassinoux, R.H. Baud and J.-R. Scherrer, in Conceptual Structures: Current Practices, Proc. of the Second International Conference on Conceptual Structures, ICCS'94, College Park, MD, August 1994, Springer-Verlag, Berlin, 1994, pp. 84–96.
- [21] G. Postma, PhD Thesis, University of Nijmegen, 1996.
- [22] B. van Bakel, PhD Thesis, University of Nijmegen, 1996.
- [23] G. Postma, B. van Bakel and G. Kateman, J. Chem. Inf. Comput. Sci., 36 (1996) 770.
- [24] G.J. Postma, B. van der Linden, J.R.M. Smits and G. Kateman, J. Chemom. Intell. Lab. Syst., 9 (1990) 65.
- [25] B. van Bakel, MA Thesis, University of Nijmegen, 1988.
- [26] Bath-meeting of full text processing, Bath, 1996.

Bas van Bakel is at the Knowledge-based Systems Group, University of Twente, Enschede, The Netherlands. Geert Postma is at the Department of Analytical Chemistry, University of Nijmegen, The Netherlands.

TrAC on the Internet

The Internet Column articles of TrAC can also be found on the Web. If you have a browser, to access the TrAC column on the Web simply point to:

<http://www.elsevier.nl/locate/trac>