USE OF AUTOMATIC KEYPHRASE GENERATION FOR CREATION OF A CONSTRUCTION THESAURUS

Automatic keyphrase generation for thesaurus construction

B. KOSOVAC
Department of Civil Engineering, University of British Columbia, Vancouver, Canada
D.J. VANIER
Institute for Research in Construction, National Research Council Canada, Ottawa, Canada

Durability of Building Materials and Components 8. (1999) *Edited by M.A. Lacasse and D.J. Vanier*. Institute for Research in Construction, Ottawa ON, K1A 0R6, Canada, pp. 2507-2516.

Abstract

The paper describes development of a thesaurus in the roofing domain. This work is part of a larger effort to investigate the potential of thesauri as an aid in product modeling. Extractor, a software module that extracts keyphrases from documents, was used for collecting candidate thesaurus terms from Internet sources. The principal advantage of the Internet as a source of candidate terms is that it reflects colloquial language: -- the language that is actually used by building practitioners and that it covers the widest range of different 'user views' on the domain. The advantage of using Extractor or similar software is that it allows processing huge text *corpora* available on the Internet and it eliminates irrelevant terms. The methodology used was found to be highly useful, although it was not sufficient by itself for constructing a construction thesaurus, as considerable human intervention was required. Though limited time resources did not allow full exploitation of Extractor's capabilities, some possibilities for customization of the software and for partial automation of a thesaurus construction process are suggested.

Keywords: Thesaurus, Internet, automatic indexing software, thesaurus construction.

1 Introduction

A thesaurus is a set of terms that are used in a specific domain of knowledge, "formally organized so that a priori relationships between concepts are made explicit" (Aitchinson 1987). Originally intended for indexing and retrieving documents, thesauri have increasingly been seen as knowledge bases and used beyond the domain of librarianship (Kosovac 1998). The overall objective of this research direction is to investigate the potential of thesauri to assist the development of product models in the

construction industry. It is believed by some (Vanier, 1994) that thesauri can assist to eliminate many semantic problems hindering the creation of robust product models for the industry. The main purpose of this research project is to explore the use of the Internet as a source of candidate thesaurus terms.

Extractor (Extractor 1998) is a machine-learning-based software module, developed by the Interactive Information Group of the National Research Council Canada, that scans an electronic document and extracts keyphrases best describing the document's "aboutness." This paper reports on the use of Extractor 2.0 as a support tool for collecting and selecting candidate terms for a thesaurus in the roofing domain, and more specifically, for low slope roofs.

In the process described, Extractor was used for a specific task and under given circumstances. Technological constraints — limited processing power, lack of programming resources for customizing Extractor and integrating it with other applications, as well as the unavailability of adequate software for advanced manipulation and analysis of Extractor's output, did not allow full exploitation of Extractor's capabilities. These constraints also precluded testing *corpora* large enough to provide statistically valid results. Therefore, the work described is of explorative nature and cannot be considered as a study that evaluates performance of Extractor 2.0. However, the patterns noticed in the analysis of the results can point to possible use of the software for related purposes and suggest possibilities for further research and/or development.

1.1 Description of the problem

The goal of the work described is to build a thesaurus based on Internet/intranet resources in the low slope roofing domain. It was also decided that it should follow the format of the *Canadian Thesaurus of Construction Science and Technology (TC/CS)*, a thorough, comprehensive, yet "dated" construction thesaurus (TC/CS 1978). The goal can be achieved as follows:

- selection of terms that can be extracted from the TC/CS to form a sectorial thesaurus,
- updating the sectorial thesaurus according to the development of the field and its terminology, and
- developing a micro-hierarchy of narrower concepts if required.

Alternatively, using the "bottom up" approach, the goal can be achieved by:

- collecting terms relevant to the field,
- selecting terms to be included in the thesaurus according to its intended purpose,
- checking the terms against the existing TC/CS thesaurus,
- organizing the terms into hierarchies following the TC/CS guidelines.

Standard sources for the initial collection of thesaurus terms usually include:

- terminological sources in standardized form; existing thesauri, dictionaries, glossaries, classification schedules, encyclopaedias, lexicons, journal indices, back-of-the-book indices, term lists, treatises on terminology of a subject field;

- literature scanning;
- question scanning;
- users', subject experts', and compilers' knowledge (Aitchinson 1987).

It must be noted that literature and question scanning play a crucial role for the usability of a thesaurus (literary and user warrant), while the other sources serve mostly for clarifying the meanings of terms, facilitating their arrangement into hierarchies and filling gaps.

1.2 Use of information technology in thesaurus construction

Along with their use for thesaurus-management, computers have been used for a long time for collecting thesaurus terms (Gilchrist 1971, Lancaster 1986, Aitchinson 1987). Their main use has been to derive terms from machine-readable sources and rank them by frequency of occurrence. Despite some successful efforts in automatic establishment of inter-term relationships, computers are still used only as support tools in this part of the process. Computers can assist human compilers by producing co-occurrence tables pointing to possible relations between terms, and clustering terms containing the same word or stem thereof.

2 Process

2.1 Approach to the problem

The TC/CS is a huge thesaurus (approximately 15,000 terms) covering a wide subject field. Searching it for all terms relevant to the subdomain would be an expert-labour-intensive process of following links throughout different hierarchies and numerous general terms, and deciding which terms are relevant and current. Possibilities for automating this task are minimal. Another problem is that the TC/CS is rather outdated, especially having in mind the significant changes in the field of low slope roofing that occurred in and around the 1980s, with the introduction of new materials and types of roofing systems.

On the other hand, the TC/CS has a thoroughly elaborated structure and well defined inter-term relationships that facilitates addition of new terms, given that their exact meaning is known. Furthermore, it is available in electronic form on the World Wide Web (http://www.nrc.ca/irc/thesaurus) thus allowing easy searching for known terms. For all these reasons using the "bottom up" approach suggested earlier seemed to be a logical solution to meet the goal of the work.

As the proposed thesaurus is intended for indexing Internet/intranet sources, the most useful source of terms would be *corpora* available on the Internet. The Extractor 2.0 documentation pointed to the suitability of the software for performing "literature scanning" of Internet sources as it:

- integrates HTML and e-mail filters and
- permits processing large *corpora* by extracting only relevant terms.

2.2 Preparatory tests

Since Version 2.0 of Extractor had been recently released (Version 3.3 is the current version), and since all known automatic literature scanning involved either title and abstract, or full-text scanning, the development of the methodology required some initial tests that would roughly examine available sources and Extractor's behaviour. The tests were done on documents retrieved by general search services such as Altavista (http://www.altavista.com) and the services listed below — and on documents from selected *web sites*. The documents were processed by Extractor 2.0 varying its *Beta* parameter; a performance measure based on the relationship between recall and precision. Here, precision means the percentage of the extracted keyphrases that are relevant, while recall represents the ratio of the number of relevant keyphrases extracted to the total number of relevant keyphrases in the document. The output of Extractor was intellectually analysed. The results and the limited resources necessitated the following modifications to the initially considered strategy:

> The query formulated with the intention to test harvesting *corpora* using automatically generated queries that combine synonyms of the top term (*summum genus*) and its immediate narrower terms;
>
> *("low-slope roof*" OR "flat roof*") AND ("built up" OR BUR OR "multi ply" OR "single ply")*
>
> did not provide optimum recall and precision within some search services. As processing a sample large enough to compensate for the deficiencies was unrealistic, the query was modified into:
>
> *("flat roof*" OR "low slope") AND ("built up" OR BUR OR roofing OR membrane*)*
>
> that better reflected the language of relevant documents. Although the new query did not eliminate all the "noise", or did it ensure absolute recall, the retrieved sets of documents seemed acceptable for the purpose.

Extractor 2.0 proved to be significantly better suited for this purpose than the previous release. The initially considered strategy was to process documents with the lowest *Beta*, identify terms that should be added as stop phrases, cluster documents based on the rest of key-phrases, process them with *Beta*=2.0, and repeat the process until the desired level of specificity is achieved. However, the inability to frequently customize the software by adding new stop phrases and the labourious task of processing the same documents more than once made this strategy unfeasible. Using the new release with the highest *Beta* (i.e. maximized recall) and simply removing the most frequent keyphrases proved to provide satisfactory results. It must be noted that the Application Program Interface (API) for the Extractor Dynamically Linked Library (DLL) allows easy integration of the software with other applications. The limited time resources mentioned earlier, however, didn't permit the use of this feature.

It was observed that long documents that tended to be of high quality and that abound with very specific terms gave only too general terms in the Extractor output. Although keyphrases derived by Extractor reflect well the subject of the documents, they did not include specific terms that would be more useful for the purpose. The efforts to semi-automatically divide documents searching for heading tags did not prove feasible with most of the Web documents. It was, therefore, done only on a small number of scholarly papers that tended to be well-organized and had a better HTML structure.

2.3 Methodology

The goal of the work is to extract relevant terms from Internet documents, and not to evaluate Extractor 2.0. However, Extractor's performance has been continuously evaluated after each step, where not possible statistically then at least intellectually, in order to consider re-design of methodology or even give up the use of Extractor 2.0 if it would not produce the desired output.

Though most of the tasks had to be performed manually the methodology tried to follow computer logic in order to examine possibilities of automating the process or at least the use of clerical instead of expert labour ("artificial dumbness" approach).

2.3.1 Corpora

The following collections have been selected:

1. Documents retrieved by general search services
   The first (15) documents retrieved by advanced search with each of the five major general search services; AltaVista, Excite, HotBot, Infoseek, and Lycos; 75 documents altogether were used. Although most of the search services perform better with other search options, for consistency, the Boolean query or the option closest to it was used in each:
   ("flat roof" OR low-slope) AND (built-up OR BUR OR roofing OR membrane*)*
   The services were searched in alphabetical order, taking care to avoid duplicates. Where relevant documents from a certain site were grouped together, only the first one was used in order to avoid language of one author and frequent appearance of the same corporate names and trademarks.
2. IRC *Roofing Resources* (Roofing Resources 1998) collection
   This collection was included as it represents the core of the collection to be searched by the future thesaurus. Files bigger than 20 KB (arbitrarily established limit) were divided by headings to form 40 documents for the extraction of keyphrases.
3. Relevant documents retrieved at *FacilitiesNet* (FacilitiesNet 1998)
   The criterion for the inclusion of this site was its high content of documents relevant to the facilities-management aspect of flat roofs that is neglected in the majority of web documents, but important to the wider context of this work. Sixty-one relevant documents were retrieved and processed.
4. Collection of selected articles
   This collection was compiled by following links from various lists of relevant

sources but without aspirations to be comprehensive, exhaustive nor of highest quality. It has been included to allow comparison of the results with those achieved by automatic harvesting of sources using general search services.

Newsgroups were not processed separately because of difficulties in accessing "Dejanews" at the time and unavailability of adequate newsgroups archives. A small number of this type of documents was however included in AltaVista hits.

2.3.2 Procedure

Each document was processed by Extractor 2.0, with its *Beta* parameter set to 2.0, meaning the maximum recall of keyphrases. The extracted keyphrases were gathered in a list. The following information has been kept for each term:

- position in the Extractor list
- relevance factor number (provided by Extractor)
- document from which extracted

and for each document:

- collection ID
- size of the file.

After processing each set of documents the results were analyzed, compared with other sets, and the sets were finally integrated and processed together. The final set of keyphrases has been compared to the TC/CS, to the existing glossaries, and intellectually analyzed.

2.3.3 Criteria

Single- and multiple-word terms were not treated separately, as the list was not too big. Singular and plural forms of the same term were counted together but terms that may have different meanings when used in plural were specially marked [PL]. The terms were first searched in TC/CS for exact matches [=]. Qualifiers from the thesaurus were ignored in this step. Terms identified as general terms in the thesaurus were additionally marked [GT] and so were those that had further developed hierarchy of narrower terms [+].

After the identification of exact matches, single-word terms from the list were also searched for occurrence in phrases from the thesaurus [PH]. Extracted terms that are used as qualifiers in the thesaurus were also identified [Q]. The rest of the terms were searched for close matches [~], meaning those having the same stem. Intellectual analysis of the remaining terms identified phrases ill defined by Extractor [*], acronyms [A], proper names [N] and also the matches that have the same form but different meaning in the thesaurus [$].

3 Results

Terms with extremely high frequency of occurrence that should have been made stop-phrases were identified early in the process. These terms were the same for each set of documents and could have also been identified by processing the list with Extractor 2.0 with the lowest *Beta* (meaning minimum recall/maximum precision).

The results from the general search services were then compared with those from selected collections. There were no significant differences noted in terms of relevance of extracted keywords that would justify laborious and intellectually demanding task of searching, evaluating, and selecting sources. The quantity and diversity of documents that can be easily retrieved by general search services can compensate for the quality of sources. As the first 20 documents from the compiled collection did not bring new terms to the list, this collection was not further processed and it is not included in the final results. However, the documents have been saved for later comparison of scholarly and natural language terms and exploration of specific subareas.

The final list consists of 1054 terms (2423 occurrences) extracted from 176 documents. Almost half of the terms extracted were single words (49 %). They accounted for almost all of the top 4% most frequent terms with only two exceptions that would normally be included in stop-phrases. The usual practice of treating such terms separately was not followed as most of the terms were also identified as single word terms in the TC/CS.

A huge number of single occurrences of a term (78%) can be explained by the insufficient size of the sample. In order to extract relevant terms from this group, they were searched for component words and stems that could also be found in other terms. Terms found in this way were ranked higher in the list as more relevant. Rough scanning of the remaining single-occurrence terms found very few terms relevant to the field.

3.1 Comparison with the existing thesaurus

Checking the final list of terms against the TC/CS reveals a high relevance of the terms extracted by Extractor 2.0 to the field. Only terms that occurred more than once were actually searched in the thesaurus and numerically processed. Among 130 terms that appeared more than twice, 56% had exact matches in the thesaurus, 12% were found only in phrases, and 6% were marked as close matches. The matches were found in all semantic classes and in various hierarchies, showing a broad coverage of the domain. Among the remaining terms in the group, 3% were proper names, 2% acronyms, and 5% were marked as ill defined meaning that they could not stand alone as meaningful terms in the thesaurus (e.g. *install, installing, single, reinforcing, requiring, flat*). Terms that occurred twice had even more exact matches (63%) in the thesaurus but were less frequently found in phrases as they included more multi-word terms. Single-occurrence terms were not matched to the thesaurus at this point. The majority of mismatches, however, do not indicate irrelevance of the terms but more often the outdatedness of the TC/CS. The frequent occurrence of the term “membranes” for example, and the phrases containing it that are not found in TC/CS reflects changes in the field of low slope roofing and its terminology. The noise-making terms come from specific kinds of documents, mostly glossaries and

book catalogues. Such documents can be easily excluded from the beginning by modifying the initial query.

Completeness of coverage is another problem that can be evaluated only after organizing terms into hierarchies (Petersen 1990). The relatively high coincidence of terms may also indicate a lack of more specific terms that would be required for developing a microthesaurus. Whether this is the case, can be established only later in the process.

3.2 Comparison with the glossary

The terms were also matched against relevant (i.e. low slope roofs related) terms extracted from a roofing glossary (Biegel 1989). The coincidence was much lower; only 31 % of the glossary terms were found in the Extractor's output. The unmatched terms were mostly very specific terms that are not a likely document topic (e.g. *alligatoring, back-nailing,* or *cutoff)*.

4 Conclusions and recommendations

4.1 Observations on performance of Extractor 2.0

Extractor 2.0 appears to be a suitable tool for collecting thesaurus terms from the Internet. Although in the work described its use required extensive manual work, it is estimated to be more efficient than both manual and automatic full-text literature scanning. The principal advantage is that it allows scanning and handling a significantly bigger number of documents, thus providing better coverage of the field and its terminology. It can be rightfully expected that the use of the Extractor's API can make the task considerably easier and can multiply the benefits of this method.

The number of phrases marked as "ill defined" was 1% of all the terms extracted. Since the total automation of the thesaurus constructing process is not considered and since ill-defined phrases cannot cause serious consequences, this percentage can be considered ignorable. Therefore increasing recall even above *Beta*=2 in order to retrieve more specific terms would probably be safe. In most cases the lack of more specific terms in the Extractor output will not represent a deficiency; very specific terms are rarely included in a thesaurus and their presence might make one of the most important decisions in thesaurus constructing — where to stop, even more difficult. However, if constructing a microthesaurus, or if for any other reason more specific terms are needed, these terms may be obtained processing larger *corpora* or narrowing searches for Internet documents to be processed.

4.2 Observations on the Internet as a source of thesaurus terms

The Internet represents an extremely useful source of thesaurus terms. It provides huge *corpora* covering numerous aspects of a domain and different vocabularies — from the highly scholarly to the most informal. Internet documents reflect the language that is current, actually used in the field, and most likely to be used in queries. The results also showed that documents randomly harvested using general search services could provide equally valuable terms as controlled subject collections. The collected documents can be further analysed to complement the

results of the described methodology.

4.3 Implications for the further work on the thesaurus

After establishing relationships between terms and organizing them into hierarchies of the existing thesaurus it can be expected that certain areas would need further development. Upon identification of such areas, gaps will first be filled with terms:

- derived by Extractor 2.0 from new sets of documents retrieved by more specific terms
- manually extracted from documents already retrieved and judged as relevant to the subfield according to keyphrases derived by Extractor 2.0.

Use of these two sources is prioritized for the reasons listed in subsection 4.2. However, these sources cannot ensure comprehensive coverage of the domain. Therefore, manual extraction of terms from alternative sources will probably be needed. Encyclopaedic and textbook type documents, roofing manuals, various kinds of term lists, and architectural details' labels are expected to best serve the purpose.

4.4 Possibilities for automating the process

Some of the tasks that could be fully or partly automated in applications used for similar purposes would include:

- automatic retrieval of relevant documents from the Internet and their processing with Extractor
- periodical processing of the list by Extractor for finding terms with significantly high occurrence and making them stop-phrases
- ranking terms by frequency of occurrence
- exclusion of terms that occur only once in large *corpora*
- automatic exclusion of geographic names
- grouping of terms containing same words or stems
- grouping of terms by co-occurrence in documents
- suggesting inter-term relationships by co-occurrence and syntax.

5 Final notes

The work described was carried out in February 1998. In the meantime a 336-terms pilot thesaurus has been developed (http://www.nrc.ca/irc/thesaurus/roofing). As these terms represent only a very small portion of the domain and of the terms collected in this process, final conclusions on the usefulness of the methodology cannot be drawn yet.

Full results of the study are available from the authors.

6 Acknowledgements

The authors wish to thank the Institute for Information Technology of the NRCC for the use of Extractor 2.0, and more specifically, to thank the software's author, Dr. Peter Turney, for his support and encouragement. The authors also wish to acknowledge the Canadian Institute for Scientific and Technical Information and the Institute for Research in Construction, both of the NRCC, for their financial contributions to this research. The authors thank both Dr. Ferrers Clark and Mr. Scott Mellon from these respective organizations for their support and encouragement during the course of the research. The authors also acknowledge the work of Prof. Colin Davidson and the IF Group in the development of the TC/CS Thesaurus.

7 References

Aitchinson J. and Gilchrist A. (1987) *Thesaurus Construction,* Aslib, London.

Biegel S. (1989) Roofing materials, in *Encyclopedia of Architecture, Design, Engineering & Construction.* v. 4, American Institute of Architects, pp. 314-9.

Extractor (1998) National Research Council of Canada, Interactive Information Group, Ottawa, Canada. Available from: http://extractor.iit.nrc.ca. [Accessed December 27, 1998]

FacilitiesNet (1998) Trade Press Publishing, Milwaukee, WI, USA. Available from: http://www.facilitiesnet.com. [Accessed December 27, 1998]

Gilchrist A. (1971) *The Thesaurus in Retrieval,* Aslib, London.

Kosovac, B (1998) *Internet/Intranet and Thesauri*, Canadian Institute for Scientific and Technical Information, Internal Report, National Research Council Canada, Ottawa, Canada.

Lancaster F.W. (1986) *Vocabulary Control for Information Retrieval,* Information Resources Press, Arlington, VA, USA.

Petersen T. (1990) Developing a new thesaurus for art and architecture, *Library Trends,* Vol. 38, No. 4, 644-658.

Roofing Resources. (1998) National Research Council of Canada, Institute for Research in Construction, Ottawa, Canada. Available from: http://www.nrc.ca/irc/roofing/roofing-waissearch.html. [Accessed December 27, 1998]

TC/CS. (1978) *Canadian Thesaurus of Construction Science and Technology*, Department of Industry, Trade and Commerce, Government of Canada, Ottawa. Available from: http://www.cisti.nrc.ca/irc/thesaurus/

Vanier, D.J. (1993) Identifying concepts and relationships in building codes: Classification system approach. *International Journal of Construction Information Technology*, Vol. 1, No. 3, pp. 53-72.

Vanier, D.J. (1994) Canadian thesaurus of construction science and technology: A hypercard stack, *Proceedings of the Joint CIB Workshops on Computers and Information in Construction (Montreal, Que., Canada)*, pp. 559-564, (*CIB Proceedings*, Vol. 165).