The Web as a source of informative background knowledge

Caroline Barrière
National Research Council of Canada
283 Taché Blvd.
Gatineau (Québec), Canada, J8X 3X7
Caroline.Barriere@nrc-cnrc.gc.ca

Abstract

In this paper, we present how a tool called TerminoWeb can be used to help translators find background information on the Web about a domain, or more specifically about terms found in a text to be translated. TerminoWeb contains different modules working together to achieve such goal: (1) a Web search module specifically tuned for informative texts and glossaries where background knowledge is likely to be found, (2) a term extractor module to automatically discover important terms of a source text, (3) a query generator module to automatically launch multiple queries on the Web from a set of extracted terms. The result of these first three steps is a background knowledge corpus which can then be explored by (4) a corpus exploration module in search of definitional sentences and concordances. In this article, an in-depth example is used to provide a proof of concept of TerminoWeb's background information search and exploration capability.

1 Introduction

In this paper, we present a system which could be used by human translators, in order to rapidly get background information on the topic of a source text. This can be particularly useful in cases where the translator receives a text which is slightly outside of his area of expertise. In that kind of situation, it is not uncommon for translators to look for a handful of introductory documents in the field, and read them diagonally in order to immerse themselves in the topic and its terminology. But finding such documents and locating their most "useful" sections can be time consuming. The system proposed in this paper would assist with both of those tasks.

The system is called TerminoWeb, which was first conceived as a software environment for terminologists to help them perform thematic searches. More specifically, the software targets the monolingual discovery and understanding of the main concepts and terms of a domain. A first version, TerminoWeb 1.0, was released as a research prototype in 2006[1] with such focus.

In thematic searches, the quest for understanding a domain relies not only on the discovery of documents which are domain specific but also informative, factual, definitional in nature. These documents must have an expert-to-novice communicative purpose (Pearson 1991). Scientific articles, written by experts for experts, are not likely to contain much background or definitional knowledge, but textbooks, glossaries, or course notes would because of their communicative purpose.

At the core of TerminoWeb is the hypothesis that definitional contexts are expressed via surface linguistic patterns. Typical definitional elements of interest are hyperonymy relations ("is a kind of"), synonymy relations ("is also known as"), meronymy relations ("is composed of"), function relations ("is used for"). Given in parenthesis above are surface linguistic patterns or "knowledge patterns" (Meyer 2001) as we prefer to call them.

The study of Knowledge Patterns (KPs) is a current research topic in the Computational Terminology community (Auger and Barrière 2008). It was also explored in the 1990s by the Natural Language Processing community working on automatic construction of knowledge bases from machine readable dictionaries. Many researchers at that time had studied what they called defining formu-

[1] TerminoWeb 1.0 has been available since December 2006 at http://termino.iit.nrc.ca.

lae, looking at how dictionary definitions are made and what characterizes them. As the Web became present in everyone's life and electronic dictionaries seemed then too static compared to richer corpus information, much research went on studying the retrieval of definitional information from corpora, more particularly specialized corpora in terminological studies. Barrière (2004) presents a comparative study of the use of defining formulae in machine readable dictionaries and corpora, providing the reader with many references to earlier and important work. The constant expansion of Wikipedia[2] is challenging our view again, showing that encyclopedic information is not necessarily static in nature, reopening the door to knowledge extraction from structured sources (Nastase and Strube 2008).

Many knowledge patterns (KPs) for different semantic relations are pre-encoded in TerminoWeb so it can look for documents containing them. In fact, these pre-encoded KPs are used in two modules: the corpus building module to assign an "informative score" to each document found on the Web, and the corpus exploration module to focus on definitional sentences within the retained documents. Such sentences should actually not only contain KPs, but rather KPs in close proximity to a term of interest. This is where TerminoWeb's term extractor module is used to search through the informative corpus built for important terms of the domain. Each term can then be studied in definitional contexts, also called Knowledge-Rich Contexts (KRCs).

The search by a translator for background knowledge related to a technical or scientific article to be translated has similarities with a thematic search, but has a different starting point (a scientific source text to be translated instead of a domain given by a client). One particular module, an automatic query generator, now part of TerminoWeb 2.0[3], is of most interest for adapting TerminoWeb from a tool for thematic searches for terminologists to a tool for translators searching for background information. The query generator uses the extracted terms from the source text and makes a set of random combinations of such terms to launch multiple Web queries. Results of these queries are integrated by the corpus building module which can then be explored.

The search for definitional knowledge on the Web at large is quite a challenge and that is what TerminoWeb attempts. Certainly many free online dictionaries (monolingual or multilingual) for the general and specialized language are available on the Web. They are a good starting point to find definitional information. Tools such as Google Define[4] provide direct links to dictionaries. The role of TerminoWeb is in complement to these tools. Its purpose is to find definitional information within any documents, whether they are dictionaries or other texts. Very specialized knowledge might be difficult to find in online dictionaries. TerminoWeb searches for informative specialized knowledge and also allows the retrieval of multiple definitional contexts all at once which allows for quick browsing of the information.

The next section (section 2) focuses on how to use TerminoWeb 2.0 in a "Background Knowledge Discovery Workflow" also giving details of each module: query generation, corpus building, term extraction and corpus exploration. Section 3 briefly reviews related work. Although there is much work in computational terminology on term extraction, there is not much work on corpus construction, or on automatic search for definitional knowledge. This makes TerminoWeb quite unique. Section 4 then concludes and points to future work.

2 Steps toward finding background knowledge

We present one single example in depth. This example was randomly chosen by the author who arbitrarily took the most recent scientific publication on NRC Institute for Research in Construction publication website[5]. This provides a real example of a scientific article for which a translator could need background knowledge to help in their translation. The article is also completely outside the field of expertise of the author, so that there is no bias in analyzing the information.

[2] Collaborative online encyclopedia found at: www.wikipedia.org.

[3] TerminoWeb 2.0 is available since June 2009 at http://terminoweb.iit.nrc.ca.

[4] In Google.com, a user can type "define: unkown_word" to obtain links to dictionaries containing definitions of the unknown word.

[5] NRC Publications for the Institute for Research Construction are available at: http://irc.nrc-cnrc.gc.ca/pubs/newpubs_e.html

The different steps in the Background Knowledge Discovery Workflow are:

1) Upload a source text.
2) Perform term extraction on the source text.
3) Build a set of queries using extracted terms and launch these queries on the Web.
4) Score/rank search result documents based on an "informative measure". Top documents are kept to form a background knowledge corpus.
5) Search through the corpus for definitional information.

We revisit each step hereafter:

1. Text upload

The user can copy-paste a source text in TerminoWeb. The source text used for our example is "Sensitivity of hygrothermal analysis to uncertainty in rain data."[6]

2. Term Extraction

The term extractor will find single-word and/or multi-word terms in the document. The number of terms to be found can be set by the user, or it can be estimated automatically based on the document's length and the actual term statistics. TerminoWeb's term extractor is based on the algorithm described in Smadja (1993). It is purely statistical, based on frequencies, and it expands frequent single-word terms into multi-words terms.

Figure 1 shows a list of extracted terms from the uploaded source text. The user can inspect and provide an Accept/Reject decision. Although this step is optional (an "accept all" can be done), user validation helps to eliminate non-pertinent terms that could introduce noise in later Web searches.

3. Query generation

Although very simple, this query generation module is important, as mentioned before, for adapting TerminoWeb to translators' needs. This module takes the list of terms and makes random combinations of 2 or 3 or 4 terms. Each combination becomes a new query sent to Yahoo API[7] to obtain a set of documents.

Figure 2 shows six queries automatically generated from the list of terms. The combination "wind-driven rain" AND "moisture content" results in 10 documents, but the combination "tipping bucket" AND "hygrothermal analysis" results in an empty set.

Three important factors will impact on results:

1. Number of queries.
2. Number of terms per query.
3. Term specificity.

The first factor, number of queries, is first a trade-off between the information gain and a longer waiting period. Although queries are launched in parallel on the server (and therefore the waiting time is not linearly increasing with the number of queries), there is still a longer wait for the user if many queries are launched. Also, more queries lead to more information, but that is not our purpose, on the contrary we wish to find more targeted information. It will be important in the future to better measure the gain from more queries versus better chosen or targeted queries. In the present version of TerminoWeb, this parameter is left to the user to decide.

Factors 2 and 3 are intimately related. When multiple very specific terms are combined, the resulting set is likely to be empty (no documents found). When few general terms are used (one at the limit) the resulting set is likely to be extremely large and inappropriate (imagine results of a query "rain" or "wall").

Generality and specificity of terms is also related to term polysemy as general terms tend to be quite polysemous. A "star" in the galaxy and a movie "star" are quite different. Such term used to launch a query on the Web is likely to retrieve documents related to both domains (universe, movies).

A quick estimate of how specific or general a word or expression is can be provided by a "hit count" measure using a search engine. In our experiment, we use Yahoo Search Engine. To pro-

[6] Sensitivity of hygrothermal analysis to uncertainty in rain data, NRCC-51257, Cornick, S.M.; Dalgliesh, A.; Maref, W., April 2009, http://irc.nrc-cnrc.gc.ca/pubs/fulltext/nrcc51257/nrcc51257.pdf

[7] Yahoo! provides a Java API to the Yahoo Search Engine which can be used for research purposes. This Yahoo API is called from TerminoWeb.

vide the reader a sense of the large range from specificity to generality, we show in Figure 1 extracted terms sorted by their hit counts. The term "hygrothermal analysis" is more specific (hit counts: 802) than "rain gauge" (hit counts: 2170000) which is more specific than "wall" (hit counts: 1520000000). Even though "wall" is an important concept in this source text studying moisture content in walls, if "wall" is used as a query term in isolation, it is very unlikely to lead the user to informative knowledge useful in understanding the source text.

An interesting feature in TerminoWeb is to provide the user with term frequency filtering by specifying lower-bound and upper-bound thresholds on hit counts. Figure 3 shows TerminoWeb's user interface to perform such frequency filtering as well as specify the number of queries and number of terms per query. Default values are set if user involvement needs to be reduced.

To further address the problem of reducing the Web search space, TerminoWeb uses the notion of domain words, or mandatory words. By making the two terms "hygrothermal" and "building" mandatory (see Figure 3), they will be included in each query generated and define the focus for the other terms. This can be used in conjunction or independently of the hit count filtering.

4. Corpus building

After the queries are performed, all the resulting documents are put together to form a large corpus that can be analyzed.

The maximum number of documents would be equal to the Number of Queries * Number of documents per query, but that is an upper bound since many queries return a smaller set than what is desired, and also, there is much document overlaps in the returned sets. It is also possible that Yahoo Search Engine returns a non-empty set for a particular query, but that TerminoWeb reduces it to an empty set because of a basic minimum content filtering applied to each document[8].

In default conditions, TerminoWeb would analyze the top 100 documents returned by the Yahoo Search Engine and give to each one an "informative score" based mainly on two criteria[9]: domain specificity and expert-to-novice communicative nature. Domain specificity is measured by the presence of the accepted terms in the document. Expert-to-novice nature is measured by the presence of knowledge patterns in the document.

Based on the informative score, the top 10 documents for each query are added to the corpus[10]. Within that corpus, TerminoWeb provides a link to the original web pages to allow the user to examine each document and decide between an Accept/Reject status. This step is optional in the present process and mostly useful for thematic searches in which terminologists would like to inspect each source from which they will select terms and contexts. If this step is not performed, the user will simply "accept all" documents and perform the next step (explore documents) on a larger set of documents.

5. Corpus exploration

Definitional contexts (knowledge-rich contexts) for terms in the source text can now be explored within the corpus built. Figure 4 shows some knowledge-rich contexts for "hygrothermal". Larger contexts can be viewed (with select button) to access a paragraph (and even link to the web page) in complement to the keyword in context (KWIC) view provided.

The first few knowledge-rich contexts (Figure 4) tell the user that (a) hygrothermal properties of common buildings are thermal conductivity, equilibrium moisture content, water vapor transmission, water absorption coefficient, etc, (b) hygrothermal is a term used to characterize the temperature (thermal) and moisture (hygro) conditions particularly with respect to climate, both indoors and out, (c) hygrothermal is an adjective pertaining to heat and humidity.[11] This information

[8] There is also a limitation on document types, as TerminoWeb can only process html and text documents.

[9] There are more criteria explained in Agbago and Barrière (2006).

[10] The number of pages to analyze and the number of pages to keep can be set by the user, but are set by default respectively at 100 and 10.

[11] (a) Article from a meeting of the ASHRAE (Association of the Society of Heating Refrigeration Air Conditioning Engineers) http://www.articlearchives.com/science-technology/engineering/761782-1.html)
(b) Green Building Advisor Glossary (http://www.greenbuildingadvisor.com/glossary/8,

gives key elements to understanding "hygrothermal" in the context of buildings.

TerminoWeb makes it also possible to view all the occurrences of a term regardless of whether it occurs in a definitional context or not. In that case, the corpus exploration module behaves as a normal concordancer allowing the user to sort the contexts on the preceding or following words. In our present example, contexts sorted by the following word are useful to discover compounds such as hygrothermal conditions, criteria, design, effects, environment, interaction, load, material, measurements, or mechanisms. Knowledge of such common term combinations is also useful to translators.

3 Related Works

TerminoWeb is quite unique because it not only contains modules that are unique (such as its corpus construction module searching for informative documents), but it also provides an interesting integration of different capabilities.

Looking at its capability to manage corpora and explore them, it is close to Corpografo (Maia and Matos 2008) system which also allows (and in many languages) to explore a corpus. Although documents must be uploaded from the user as it does not provide web searches.

For query generation, our work was inspired by the work of Baroni (2004, 2006) who first suggested query combinations of common words to build a corpus of general knowledge (Baroni and Bernardini 2004) and then presented the same idea for specialized language (Baroni et al. 2006). Their searches perform no ranking based on informative scores (as this is not the purpose of their work). Although many researchers have use knowledge patterns (Auger and Barrière 2008) for discovering term relations, their use in scoring documents is unique to TerminoWeb.

A large pool of research exists in computational terminology around the problem of term extraction. Although a simple frequency based approach is part of TerminoWeb, there are more sophisticated algorithms being developed in the community (see Cabré Castellvi et al. 2001 for a review of earlier systems and Drouin 2003 for a new trend of term extraction based on comparing corpora).

Overall, although the value of "disposable corpora" for translators has been mentioned earlier (Bowker 2002, Varantola 2002), not much work has been done for finding support information for translation either in the source language to better understand the concepts or in the target language to validate the use of an equivalent (although see Sharoff et al. 2006 for an interesting use of comparable corpora).

4 Conclusions and future work

TerminoWeb is a constantly evolving prototype available on the web for terminologists, translators and other language workers in need of specialized corpora that will help them understand terms related to a domain or a source text. It is provided as a prototype to obtain feedback and stimulate ideas to pursue research in computational terminology. Version 1.0 was specifically oriented toward the task of thematic searches and version 2.0 is introducing new modules to open to the needs of translators.

The scenario we presented in this article shows the use of TerminoWeb in a particular set of steps to help translators find background information about a source text to translate. The starting point is a source article, and TerminoWeb generates a whole corpus containing background information about the topic of the article. To obtain such corpus, the user involvement is minimal, and could be limited to the filtering of the set of extracted terms to be used for the queries and to the specification of number and sizes of queries (and their mandatory terms if any).

TerminoWeb's purpose is to build a domain-specific (or source-text specific) background knowledge corpus. This goes beyond finding specific term definitions. For many terms (unless they are very specialized), definitions can be found using Google Define. In our example, such search for the word "hygrothermal" would lead to a brief Wiktionary entry "Of or pertaining to both humidity and temperature"[12]. The use of TerminoWeb is certainly not easier than such search. On the contrary, it is more involved, and can be used in complement to find more information about a topic or a

(c) Luciferous Logolepsy – a collection of 9000 obscure English words (http://www.kokogiak.com/logolepsy/ow_h.html

[12] http://en.wiktionary.org/wiki/hygrothermal

set of terms related to a source text. It aims at finding more "grounded" information, by the fact that such information is used in texts, and it will help terminologists and translators confirm their understanding of terms by viewing them not only in multiple contexts, but in contexts related to their interest (e.g. hygrothermal in a building context).

As future work, we wish to look closely at this intriguing balance between term specificity, polysemy and domain-relatedness. This should lead to a fully automatic query generator taking all parameters in consideration. Also, and mostly, we need to work closely with translators to understand better the value of our tool in their work environment.

The search for background knowledge by translators is unfortunately not something well documented. Désilets et al. 2009 perform a study where they observe translators. This type of study will help understand more the practice and the needs of translators not only for term equivalent searches, but also for background information searches, and for equivalent validation searches. Also, by providing TerminoWeb 2.0 as an online prototype, we hope to facilitate future collaboration and joint studies in which translators would provide feedback on the prototype and help have a better understanding of how to adapt or modify it to their specific needs.

References

Agbago, A and C. Barrière. 2005. Corpus Construction for Terminology. *Corpus Linguistics Conference*, Birmingham, UK, July 14-17.

Auger A. and C. Barrière. 2008. Pattern-based approaches to semantic relation extraction: A state-of-the-art. *Special Issue on Pattern-Based Approaches to Semantic Relation Extraction, Terminology*, 14(1), 1-19.

Baroni, M. and S. Bernardini. 2004. BootCaT: Bootstrapping Corpora and Terms from the Web. *Proceedings of LREC'2004*.

Baroni, M., Kilgarriff, A., Pomikalek, J. and Rychly Pavel. 2006. WebBootCaT: instant domain-specific corpora to support human translators. *Proceedings of the 11th Annual Conference of the European Association for Machine Translation, EAMT-2006*, Oslo, Norway.

Barrière, C. 2004. Knowledge-Rich Contexts Discovery. *Proceedings of the 17th Conference of the Canadian Society for Computational Studies of Intelligence on Advances in Artificial Intelligence, Canadian AI 2004*, London, Ontario, Canada, May 17-19.

Barrière C. and A. Agbago. 2006. TerminoWeb: a software environment for term study in rich contexts. *Proceedings of the International Conference on Terminology, Standardization and Technology Transfer*, Beijing, 103-113.

Bowker, L. 2002. Working Together: A Collaborative Approach to DIY Corpora. *First International Workshop on Language Resources for Translation Work and Research*, Gran Canaria, May.

Cabré Castellvi, M.T., Estopa R. and J.V. Palatresi. 2001. Automatic term detection: A review of current systems. In Bourigault D., Jacquemin C., L'Homme M.C. (eds) *Recent advances in Computational Terminology*, vol. 2, pp. 53-87.

Désilets, A., Melançon, C., Patenaude, G. and L. Brunette. 2009. How translators use tools and resources to resolve translation difficulties: an ethnographic study, *Beyond Translation Memories Workshop, MT Summit*, Ottawa.

Drouin P. 2003. Term extraction using non-technical corpora as a point of leverage. *Terminology* 9(1), pp. 99-117.

Maia, B. and S. Matos. 2008. Corpografo V.4 – Tools for Researchers and Teachers using Comparable Corpora, *Proceedings of LREC'2008*, Marrakech, Morocco.

Meyer, Ingrid. 2001. Extracting knowledge-rich contexts for terminography, in D. Bourigault, C. Jacquemin, L'Homme M.C. (eds) *Recent Advances in Computational Terminology*, chapter 14, John Benjamins.

Nastase V. and M. Strube. 2008. Decoding Wikipedia Categories for Knowledge Acquisition, *AAAI 2008*, pp. 1219-1224

Pearson, J. 1998. *Terms in Context*, John Benjamins Publishing.

Smadja, F. 1993. Retrieving collocations from text: Xtract, *Computational Linguistics*, 19(1), 134-177.

Sharoff, S., B Babych, and A. Hartley. 2006. Using comparable corpora to solve problems difficult for human translators. *Coling-ACL 2006: Proceedings of the Coling/ACL 2006 Main Conference Poster Sessions*, Sydney, pp.739-746.

Varantola, K. 2002. Disposable corpora as intelligent tools in translation, *Cadernos de Traduçao IX – Traduçao e Corpora*, Vol. 1, No 9,171-189.

TerminoWeb 2.0

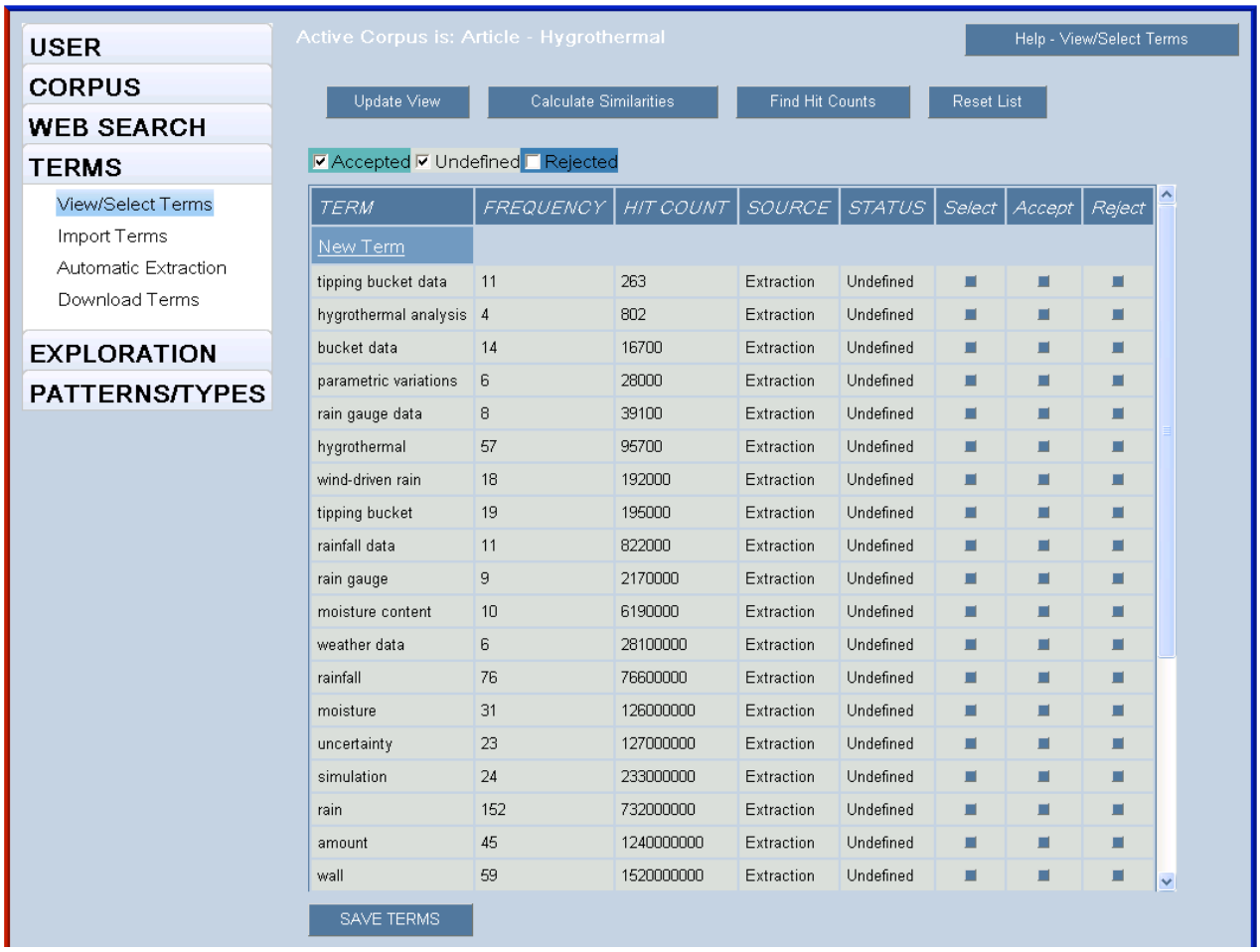

Figure 1. Terms extracted from the source text.

TerminoWeb 2.0

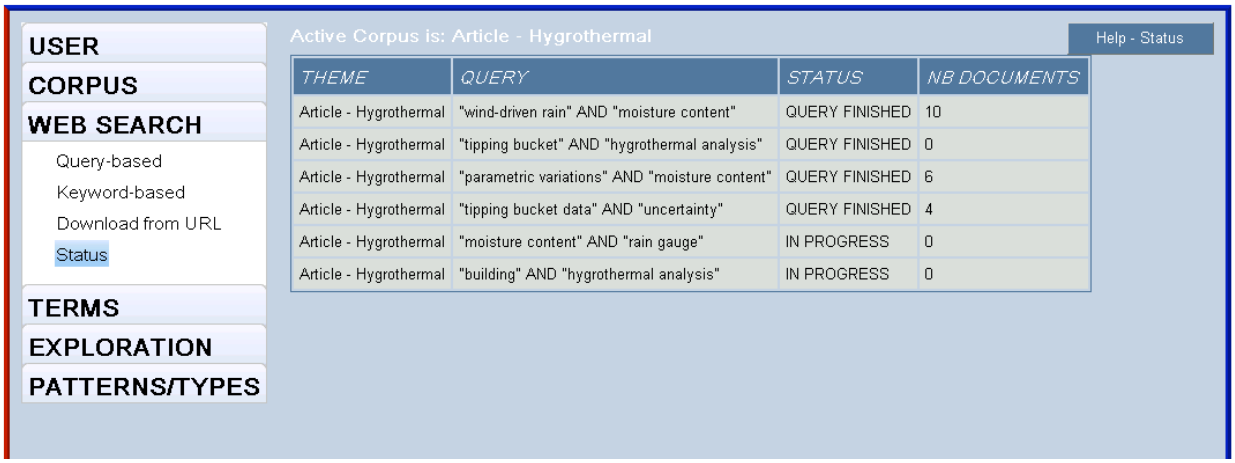

Figure 2. Automatically generated queries from extracted terms.

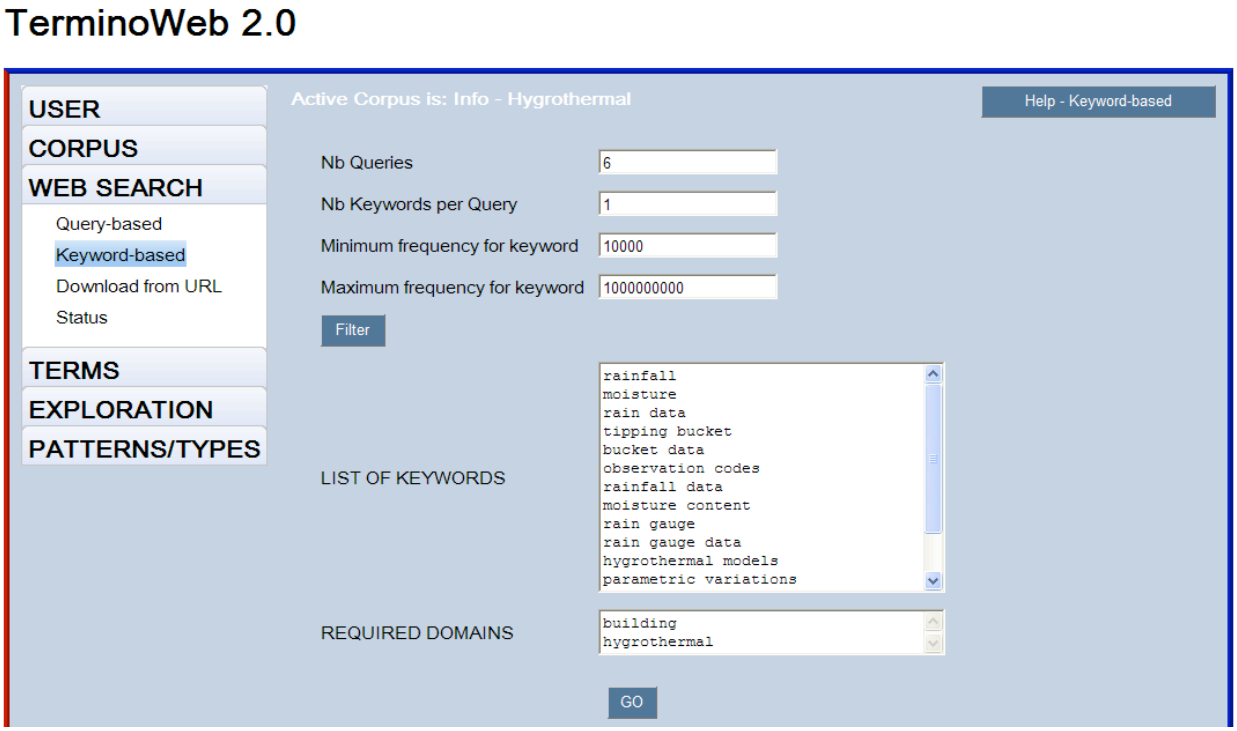

Figure 3. User Interface for the query generator.

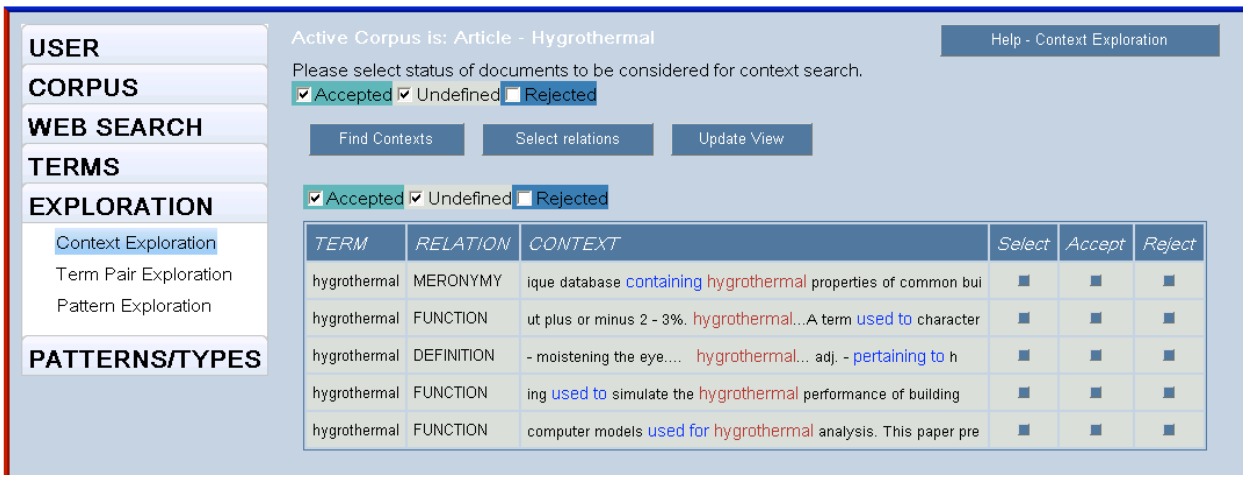

Figure 4. Knowledge Rich Contexts found for “hygrothermal”.