

Domain-Specific QA for the Construction Sector

Zhuo Zhang¹

Lyne
Da Sylva²

Colin
Davidson³

Gonzalo
Lizarralde²

Jian-Yun Nie¹

Université de Montréal

CP. 6128, succursale Centre-ville, Montréal

Québec, H3C 3J7 Canada

¹{zhangzhu,nie}@iro.umontreal.ca

²{Lyne.Da.Sylva,Gonzalo.Lizarralde}@Umontreal.CA

³dav0528@attglobal.net

ABSTRACT

Previous research on Question-Answering (QA) has focused on general domain questions. The general approach is to first recognize Named Entities (NE) in both texts and questions; then the most relevant answers (or passages) are selected (after an IR selection) according to the type of question and the NEs included in the possible answers. In this paper, we extend this general approach to domain-specific questions in the construction sector. This extension allows us to answer questions such as “What material is appropriate for ...”, which cannot be answered by a general QA system. Our approach is based on a domain-specific thesaurus, which contains a large set of domain-specific concepts organized into a hierarchy. Generic concepts such as “material” are considered as semantic categories. Our experiments on a technical corpus in construction show that this approach is effective: using our extension, we can obtain improvements on Mean Reciprocal Rank (MRR) of about 10%.

Categories and Subject Descriptors

H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval – *retrieval models, search process, selection process.*

General Terms

Algorithms, Design, Experimentation.

Keywords

Information retrieval; domain-specific thesaurus; question answering system; specialized names entities.

1. Introduction

The goal of Question-Answering (QA) is to provide a direct

answer to the question of a user, instead of a reference which may contain such an answer in the case of Information Retrieval (IR). This problem is becoming increasingly important due to the huge number of documents available on the Web. The low accuracy of the current IR systems and search engines also make some groups of professionals reluctant to use search engines in their professional activities. Professionals in the construction sector are part of them. Those professionals have their own work habits, and they consider browsing the Web as an inefficient way to find professional information. In order to meet the requirements of these professionals, two aspects of the current search engines have to be improved. 1) Search engines have to be specialized in the documents they provide. Professionals in the construction sector would like to have specialized search engines that only provide construction-related documents. 2) Search engines should provide accurate answers to their questions. This makes it possible to improve search engines with a specialized QA.

The first problem (specialized documents) can be solved by collecting documents according to a specific domain. The current techniques (such as classification) already allow us to create specialized document collections from the Web. Although every document in the collection cannot be entirely relevant to a specific domain, the fact that most of the documents irrelevant to the domain have been filtered out makes the list of retrieved answers less noisy than with the current search engines. In this paper, we will not deal with this problem. Rather, we will concentrate on the second problem which requires more research effort.

While a lot of effort has been spent on general-purpose QA, few research projects have focused on domain-specific QA, such as in the construction sector (cf. [4]). In this paper, we extend the existing approaches to QA to deal with domain-specific questions. In particular, our QA system makes heavy use of a domain-specific thesaurus, which contains a large hierarchy of concepts in the construction sector. Concepts in the hierarchy are also considered as semantic categories. Questions (in particular certain *what*-questions) then can focus on these concepts, such as in “What material ...”. An appropriate answer to this question may contain subtypes of “materials” such as “cement”, etc. The processing of these categories is inspired by that for NEs. Therefore, the general steps to process domain-specific questions with the use of a thesaurus are similar to general domain QA.

2. General-domain QA

General QA systems have initially focused on answering common sense questions. Namely, they often try to answer questions whose answer types belong to common NE types, i.e., an NE type that is domain-independent, such as date, person name, organization and so on. For example,

Question 1: When was Trec-10 held?

Question 2: Who is the President of USA?

The expected answer types for these two questions are DATE and PERSON respectively. These types are added into the questions so that the general QA system can return precise answers for this kind of question.

General approaches to QA usually follow the following schema. First, Information Extraction (IE) is performed on both documents and questions. In particular, Named Entities (NE) are recognized and tagged (using gazetteers [1], [7], heuristics [6], [5], [10], or machine learning [9]). Then an IR module is used to retrieve a relatively small set of passages which contain the keywords of the question. Finally, specific analyses on question type and the possible form of answers are used to re-rank the candidate answers.

Typical questions which previous research has addressed are certain types of WH-questions. For these questions, a key concept in the question is identified as the asking point (e.g. what temperature ...). For domain-specific QA, it is necessary to extend the existing QA approaches by adding knowledge on domain-specific concepts and questions, to handle domain-specific asking points. This is the aim of the research project described in this paper: the construction of a domain-specific QA system for the construction sector. We assume that all the documents in which we try to locate answers are related to construction. We first use an existing IR system - Okapi [8] - for the basic passage retrieval. The techniques we develop are either integrated into the Okapi indexing and search process, or used in a post-processing of the retrieval results. Our approach, described below, combines several existing methods described in the literature. A domain specific NE is indeed a semantic category of concepts that is identified in the thesaurus. We consider such categories as special types of NE, and questions can be asked on them.

For example, we will be able to deal with questions such as “what material is the most suitable to the constructions in the Northern areas of Quebec?”, in which “material” is considered as a type of special NE. Notice that for many open-domain QA, one can only ask questions on common types of NE such as “what is the date of independence of the USA?”.

3. Extensions to domain-specific QA

A professional in the construction sector may wish to ask the following question:

Question 3: “What materials are best suited for houses in the Montreal area?”

General QA systems cannot determine an expected answer type for this question and adopt a general IR search (or use general domain resources such as WordNet). To answer more specific

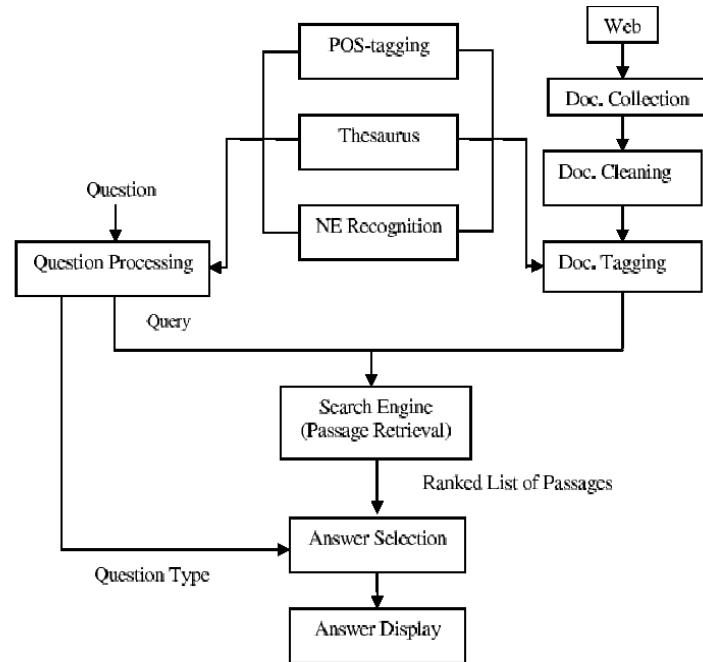


Figure 1: Workflow of the QA processes

questions than those on general NEs, one has to use more knowledge. However, world knowledge is infinite; even the largest knowledge base can only store a part of all concepts and technical terms for all domains. Our approach tries to use domain-specific knowledge, which is more available than general world knowledge.

In order to extend the general QA approach based on NEs to a specialized area, the key is to extend general NE types to specialized categories, so that questions can also be asked on the latter. Just as common NE types, specialized categories are types of (specialized) concepts.

In addition, in a specific domain, many technical terms are compound terms. Traditionally, single words are used as indexes for the first-step passage selection with an IR system. This is not precise enough. The problem of compound terms is also an important factor affecting the quality of QA systems. Thus, recognizing compound terms is an important aspect of our domain-specific QA system.

To deal with both problems, we make use of a specialized thesaurus as the backbone of our QA processing when specialized concepts are involved. This allows us to process specialized questions, in addition to other question types (such as definitions and basic NE questions, whose treatment is described below).

The global processing workflow of our QA system is shown in Figure 1.

3.1 Thesaurus

The use of a specialized thesaurus in a domain-specific QA system is essential. It helps determine the appropriate meaning of terms in the domain. For example, one of the common meanings of the term “concrete” is its adjectival meaning, akin to “real”. However in the construction domain, its meaning is “a hard strong building material”. In order to reduce ambiguity, some semantic

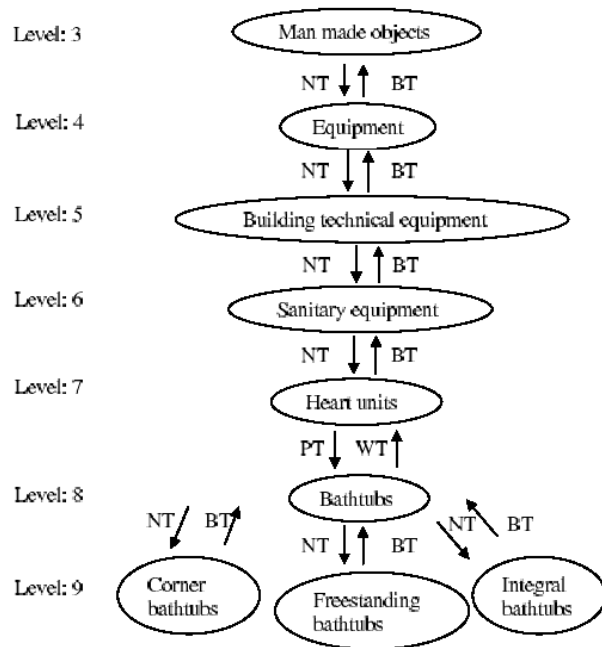


Figure 2: Excerpt from the thesaurus

information is added to this kind of term. In our processing of construction-related documents and questions, the term “concrete”, will be tagged with the semantic category “building material” which is a general concept in our thesaurus. The thesaurus that we have used is the Canadian Thesaurus of Construction Science and Technology [3]. This thesaurus is a vast network of approximately 15,354 concepts with approximately 26,000 links between them. It describes terms (concepts) and their relationships. Terms are organized into 11 levels, from 0 to 10 (from most general to most specific). The relationships defined between terms are as follows:

- UF: Used for (with its converse, US: Use)
- BT: Broader term relationship
- NT: Narrower term relationship
- WT: Whole term relationship
- PT: Part term relationship
- RT: Related term relationship
- GT: General related term relationship

Figure 2 is a small excerpt from the thesaurus (with a subset of relations illustrated).

We use the BT and NT relationship between general and specific concepts. Any concept stored in the thesaurus can be a potential semantic category on which a user may ask a question. Our system tries to assign dynamically a category for each term contained in thesaurus. Then, words in both questions and documents are tagged with categories from the thesaurus (see details below); the concept categories are also used as indexes. This constitutes an addition of semantic information. Therefore, searching is based not only on keyword search but also on concept search to some extent: the user’s query is enhanced with

categories as additional keywords. This category-based search strategy is used to make the first selection of passages.

3.2 Tagging categories in documents and questions

3.2.1 Tagging the document collection

Tagging documents relies on the semantic information provided by the thesaurus. The method used is as follows: for a term appearing in the thesaurus, we assign the direct parent of this term as its category. For the root node, its category is itself.

For example, we want to assign categories for terms “bathtubs”, “corner bathtubs”, “integral bathtubs”, and “freestanding bathtubs”. Figure 2 shows that “heart units” is the direct parent of “bathtubs”, thus, its category is “heart units”. Similarly, “bathtubs” is defined as the category for “corner bathtubs”, “integral bathtubs”, and “freestanding bathtubs”. This will allow a passage containing “corner bathtubs” to be considered as a possible answer to the question of “What bathtubs do you want to put in your bathroom?”. Thus, appropriate passages can be retrieved if they contain concepts of a level lower than words in the question. We have limited inferences of this type to a single level because extending to an arbitrary level will potentially include too much noise. Once the collection is tagged (and indexed), it is available for querying.

3.2.2 Tagging the user’s question

The user’s question is processed in the following way: the part-of-speech of each word of question is first determined, to identify the first head noun as the “identifying word” (for example, “bathtubs” above). This is a crucial indicator of the focus of the question. If the identifying word is not included in the thesaurus, the category of the identifying word is Null, and we only return the general IR results to the user without further processing. (In our experiments, this case occurs 18 times out of 100 questions.) If the head noun does occur in the thesaurus, we assign a thesaurus-based category for the identifying word (as detailed above, except that terms containing no sub-term, i.e. leaf nodes, are assigned the Null category). This constitutes the category (or type) of the question. For example,

Question 4: What bathtubs do you want to put in your bathroom?

The identifying word in Question 4 is “bathtubs”. It is not a leaf node from the structure described in Figure 2. Thus, the category of Question 4 is “bathtubs”.

3.2.3 Matching categories in documents and questions

As stated above, “bathtubs” is the category of the terms in documents such as “corner bathtubs”, “integral bathtubs”, and “freestanding bathtubs”. Then, if we submit Question 4 to Okapi, the passages that contain these terms will be regarded as containing the same category as Question 4, during the category-based search. This will broaden the coverage of the retrieval. Here, it happens that the category of the terms is contained within each; but it is obviously not always the case, hence the need for a thesaurus.

3.2.4 Tagging compound terms

Compound terms (composed of two or more single words) are also identified with the help of the thesaurus. For example, “heart units” is identified as a specific term in the construction area, instead of two separate words “heart” and “units”; it is associated with the category “sanitary_equipment” (an underscore is added to accommodate word-based searching).

3.2.5 NE recognition

The document collection is preprocessed to tag as NEs all proper nouns appearing in our gazetteers and all recognizable patterns (eg. dates). The question is also tagged in the same way. In addition, common nouns in the question which are types of NEs are tagged as well (for example, the nouns “company” and “organization” are tagged as ORGANIZATION).

3.3 Search strategy

A general IR system only uses keywords. Our domain-specific IR system also uses the semantic categories tagged in both documents and questions. For example, for a question such as “what material ...”, a passage containing not “material” but “wood” can still be a possible answer. Therefore, our IR search integrates the tagged semantic categories and NE types as part of the query.

The returned list of 50 paragraphs (with their initial ranking) will be submitted to the following post-processing according to the type of the question.

3.3.1 Question type

We identified four question types: Definition, Named Entity, Category and Keyword question types. A definition question refers to a question about the meaning of a concept. The typical form is “what is ...” (e.g. What is corrosion?). A NE question refers to a question whose answer is a NE (e.g. When ..., Who..., Where ...). Both of these types of questions have been studied in previous research on QA. We use similar approaches in our system. A category question refers to a question that looks for a domain-specific concept, such as “What material is ...”. This type of question is specific to the application area. The last type of questions – a keyword question – refers to questions for which no particular type of answer has been determined. In that case, we only provide the IR result without any post-processing.

The following focuses only on NE and category questions.

The IR system we used is Okapi. It is used to identify the first 50 most relevant passages (paragraphs) using a set of keywords. These 50 passages are then submitted to a post-processing for a further selection of the appropriate answer.

3.3.2 Answer selection

In NE search, a paragraph is retained (or re-ranked higher) if it meets the following criteria: it contains a sentence which provides the required type of NE and which is connected closely with the other NEs or keywords of the question.

The same strategy is extended to questions on specific semantic categories: We require an answer paragraph to contain a sentence that provides a concept in the specific semantic category, in connection with the other concepts and keywords specified in the question.

For both NE and semantic category questions, an empirical calculation of the weight of the paragraph is defined, which combines the weight produced by Okapi (based on keywords) and a weight determined according to whether the required NE or concept appears in a sentence, as well as the number of the concepts of the question appearing in the same sentence. The final weight of a paragraph takes the form of:

$$w(p, q) = \alpha w_{kw}(p, q) + \beta w_{Sem}(p, q)$$

where $w_{kw}(p, q)$ is the weight of paragraph p for the question q according to keywords, and $w_{Sem}(p, q)$ is the weight assigned according to NEs and semantic categories: if the required NE or semantic category appears in the paragraph, this weight is determined as the proportion of the number of occurrences of all the NEs and tagged semantic categories of the question in the paragraph, in comparison with the total number of words in the question. α and β are parameters tuned empirically. In our current setting, $\alpha = 0.3$ and $\beta = 0.7$.

We show here some examples of processed questions.

Question 5: What organization in Canada is in charge of registering earthquakes and seismic activity?

Keywords: organization, Canada, charge, register, earthquakes, seismic, activity

Compound terms: Null¹

NE type: ORGANIZATION

For this question, the following query is sent to Okapi: organization, Canada, charge, register, earthquakes, seismic, activity, ORGANIZATION. The question type is NE, of type ORGANIZATION.

Question 6: What are the common thermoset foams used in frame construction?

Keywords: thermoset, foams, frame, construction

Compound terms: frame_construction

Category type: product_forms²

For this question, the following query is sent to Okapi: thermoset, foams, frame, construction, frame_construction, product_forms. The type of this question is “Category”, and the required category is “product_forms”, which is the category for “foams”.

4. Experiments

Our experiments have shown an improvement in ranking relevant answers.

4.1 Document collection and question set

The document collection contains 240 articles from the Canadian Building Digest published between 1960 and 1990 by NRC's Institute for Research in Construction and its predecessor, the Division of Building Research. The size of this collection is about 8M bytes. The topics reflect the diversity of the industry and cover virtually every aspect of design and construction in Canada (<http://irc.nrc-cnrc.gc.ca/cbd/cbd-e.htm>). One hundred (100) test questions have been defined, which reflect the way that

¹ Indeed, “seismic activity” does not belong to the thesaurus.

² In the thesaurus, “foam” is a NT of “product form”.

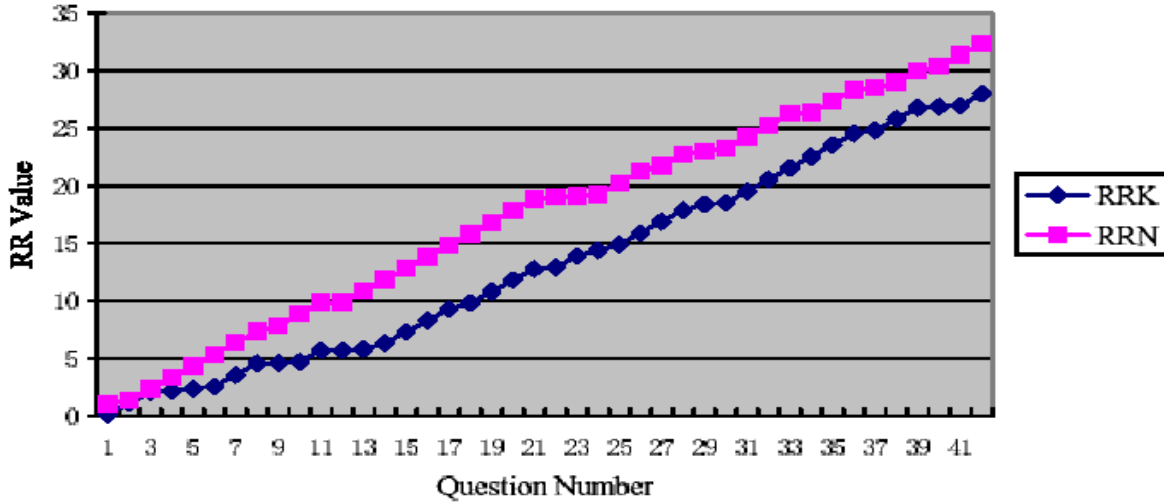


Figure 3: Keyword-based search vs. extended keyword-based NE search

professionals in construction may ask questions (the test questions are based on actual, expert questions, but may have been reduced somewhat to remove extraneous information). One passage (paragraph) in the collection is identified as the correct answer to each question. This pairing has been done manually by domain experts. The composition of these questions is as follows: 42% of them are Named Entity questions (e.g., “What is the address of the Educational Facilities Laboratories Inc.”), 40% belongs to Category questions (e.g. “What product is used to remove the stains caused by water?”), and the remaining 18% do not belong to these two types, and they are Keyword questions.

4.2 Evaluation method

We use mean reciprocal answer rank (MRR) and reciprocal answer rank (RR) to measure QA performance. The calculation formulas for MRR and RR are as follows:

$$MRR = \frac{1}{N} \sum_{i=1}^N \frac{1}{rank_i}, \quad RR = \sum_{i=1}^N \frac{1}{rank_i}$$

where N represents the number of questions in the test set; $rank_i$ represents the rank of the i -th question's correct answer. If the correct answer is not included in the list of answers, its rank is set at infinite.

4.3 Search using NEs

Figure 3 shows the performance (RR) for keyword search and with a post-processing on NEs. In this experiment, we only consider the 42 questions which have been identified to involve NEs and semantic categories.

The MRR value of Keyword search for the 42 questions is 0.6663, while the MRR value of NE and semantic category search is 0.7698, which represents an improvement of 10.35%. In

particular, we observe improvement for 33.33% of questions, no change for 47.62% of the questions (however, 95% of them already have the correct answer in first position, while the correct answer for 5% other questions does not appear in the answer list of Okapi), and degradation for 19.05% of questions. Globally, we can see that the specific processing on NE helps to improve the quality of QA. This conclusion is similar to that of previous research.

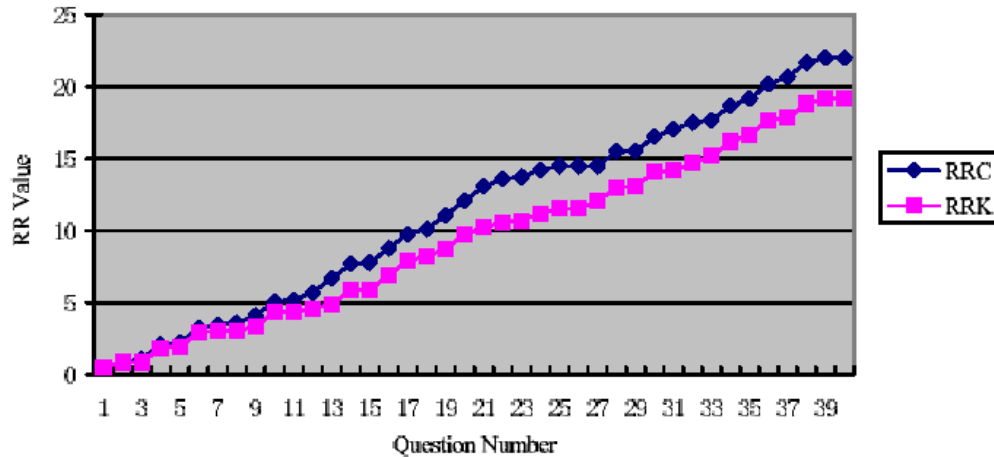
4.4 Search using semantic categories

For 40 questions, semantic categories are involved. Figure 4 shows the RR for these questions using keyword-based search and the search considering semantic categories, respectively.

The MRR value of keyword search is 0.4789, whereas the MRR for Category search is 0.55. This represents an improvement of 7.11%. In particular, the performance for 35% of the questions has been improved using category search. The performance for 55% of questions is unchanged, and the performance for 10% of questions has decreased. This result confirms that a specific consideration of semantic categories can further improve the quality of answers to domain-specific questions. Recall that if this processing were not performed, one would have used only the keyword-based search for these questions. Therefore, the consideration of domain-specific questions is important for our application area.

The remaining 18 questions out of 100 do not contain either semantic categories or NEs. They are simply evaluated by Keyword search.

Globally, for all the 100 questions, the MRR value of Keyword search using Okapi is 0.5826. The MRR value of the search with post-processing is 0.6545. The absolute improvement in the performance is 7.19% (or a relative improvement of 12.34%). If we ignore the 18 questions on which the post-processing search has no effect, the improvement of the performance is 8.77%. This



RRC: Reciprocal answer Rank for Category search
RRK: Reciprocal answer Rank for Keyword search

Figure 4: Keyword-based search v.s. search based on semantic categories

result is very encouraging. It shows that our post-processing, although still simple, is quite effective.

5. Conclusions

Our main objective is to significantly lessen the burden associated with information search, well known to be a hampering factor in the development of the construction industry [2]. The decision to adopt a QA system rather than a document retrieval system is motivated by this. And in this, we have shown that using domain knowledge, as expected, improves system performance.

One limitation of the existing thesaurus is its scope: various words and phrases used by potential users of our system do not occur in the thesaurus; this is due to a number of factors, one being the thesaurus's age, and another, its avowedly "normative" nature. Only approved terminology, by the standards of the Division of Building Research officials, made it into the thesaurus during its development in the 1970s. This of course reduces the potential for spotting appropriate answers in general documents. When the system's database includes documents on the Web, this may prove very damaging. Also, the thesaurus contains only nouns, whereas some verbs may be useful for searching. Ongoing efforts aim at broadening the thesaurus' coverage and at incorporating verbs and adjectives, as pseudo terms linked to existing thesaurus terms, to improve retrieval.

Apart from the thesaurus, a remaining hurdle is the potential information overload associated with a long list of answers, some of which may be duplicates. To reduce the user's time spent in assessing the answers, we are in the process of adding another step of post processing, that of clustering the answers.

6. Acknowledgments

We wish to express our special thanks to the Laboratoires universitaires Bell and the Natural Sciences and Engineering Research Council for the group grants given to C.H. Davidson,

enabling this research to be performed. We also thank our additional research staff (in alphabetical order): N. El Khoury, F. Jin, M. Léger-Rousseau, L. Liu, L. Moulet, M. Ngendahayo, L. Shi and Q. Zhang.

7. References

- [1] O. Bender, F.J. Och, H. Ney, Maximum entropy models for named entity recognition, *Proc. of CoNLL shared task contributions*, Edmonton, 2003.
- [2] C.H. Davidson, Technology Watch in the Construction Sector : Why and How ?, *Building Research and Information*, Vol. 29, No 3, 2001:233-241.
- [3] C.H. Davidson, M. Julien, P. Garneau, J.-J. Chailloux, *The Canadian thesaurus of construction science and technology*, <http://irc-nrc-cnrc.gc.ca/thesaurus/>, 1995 (last visited Jun. 7 2004).
- [4] A.R Diekema., O. Yilmazel, J. Chen, S. Harwell, L. He, E.D. Liddy, What do You Mean? Finding Answers to Complex Questions, *Proc. of the AAAI Spring Symposium: New Directions in Question Answering*, Palo Alto, 2003: 87-93.
- [5] G. Eriksson, K. Franzen, F. Olsson, L. Asker, P. Liden, Using heuristics, syntax and a local dynamic dictionary for protein name tagging, *Proc. of Human Language Technology*, San Diego, 2002.
- [6] Sheffield NLP group, *Named entity recognition from diverse text types*, <http://gate.ac.uk/>
- [7] A. McCallum, W. Li, Early results for named entity recognition with conditional random fields, feature induction and Web-enhanced lexicons, *Proc. of CoNLL shared task contributions*, Edmonton, 2003.

- [8] Stephen E. Robertson, Steve Walker, Micheline Hancock-Beaulieu, Aaron Gull, Marianna Lau: Okapi at TREC. *TREC 1992*: 21-30.
- [9] E.F.T.K. Sang, F.D. Meulder, Introduction to the CoNLL-2003 shared task: Language independent named entity recognition, *Proc. of CoNLL shared task contributions*, Edmonton, 2003.
- [10] D. Wu, G. Ngai, M. Carpuat, A stacked, voted, stacked model for named entity recognition, *Proc. of CoNLL shared task contributions*, Edmonton, 2003.