APPLYING CASE-BASED REASONING TO EMAIL RESPONSE

Luc Lamontagne and Guy Lapalme
Département d'informatique et de recherche opérationnelle, Université de Montréal,
CP 6128 succ. Centre-Ville, Montréal, QC, H3C 3J7, Canada
{lamontal,lapalme}@iro.umontreal.ca

Keywords: Email Response Systems, Case-Based Reasoning, Textual Cases.

Abstract: In this paper, we describe a case-based reasoning approach for the semi-automatic generation of responses to email messages. This task poses some challenges from a case-based reasoning perspective especially to the precision of the retrieval phase and the adaptation of textual cases. We are currently developing an application for the Investor relations domain. This paper discusses how some of the particularities of the domain corpus, like the presence of multiple requests in the incoming email messages, can be addressed by the insertion of natural language processing techniques in different phases of the reasoning cycle.

1. INTRODUCTION

It is expected that more than six (6) trillion electronic messages will be exchanged this year. With over 20% of this volume being exchanged for commercial purposes, it reflects an increasing usage of this communication mode within enterprises. A recent survey[1] indicates that more than 70% of the enterprises deem electronic mail either important or very important for their marketing strategy. However, only 10% of the enterprises[2] are prepared to adequately manage the volume of messages resulting from the interactions with their customers. The insertion of information technology is hence foreseen as a mean to confront this increased demand while maintaining the quality of the response.

The Mercure project at the University of Montreal aims to study different architectures for the automatic response of email messages. During the first phase of the project, some experiments highlighted the potential and limitations of a combination of some Natural Language Processing (NLP) techniques, information extraction and text generation, to address such a task [*Kosseim et al.*]. In the second phase of the project, a combination of classification, question-answering and case-based reasoning were selected as candidate techniques. Classification techniques help in the routing of messages. Question-answering techniques find factual information from a collection of documents made available at a domain-specific web site (in our case, the Investor Relations domain). And, as reported in this paper, we are also investigating the application of textual case-based reasoning (CBR) techniques to generate responses to incoming email messages. This CBR module exploits a corpus of email messages comprising requests from investors and responses from financial analysts. This provides the basis for constructing the case base and the other knowledge containers of the system.

Our motivations behind this research are three-fold. First, from a commercial point of view, while this field is expected to grow in 2002 to $210 billions in revenue, only 10% of this potential market has been penetrated. Second, from a technical point of view, the management of electronic mail messages ideally requires systems that can combine some text understanding and generation functionalities. Since the robustness of the current NLP techniques is not sufficient to address such problems, other directions must be considered at the time being. Third, case-based reasoning is, by its nature and that of the problem we seek to solve, one of the most promising approaches for this task. The design of a CBR email response system relies on a corpus of antecedent messages, a resource that is representative of the domain of discourse and of the various problems tackled during email exchanges. Furthermore, the "search and

[1] By Forrester Research.
[2] By Gartner Group [Gar00]

adapt" reasoning scheme of a CBR system offers a natural mapping to the two phases of email response, *i.e.* the analysis of incoming requests and the synthesis of relevant responses. Recent work grouped under the "textual CBR" banner has proposed extensions to CBR systems to reason with experiences contained in textual documents. While these provide an interesting basis, our research aims to overcome limitations related to the precision of the retrieval and to the adaptation of the textual solution (response).

This paper discusses various aspects of the email response CBR module we are developing for the Investor Relations domain. In the next section, we discuss some of the properties of our domain corpus and illustrate how it is exploited by the proposed CBR module. The following sections describe the main issues pertaining to the authoring, the retrieval and the adaptation of antecedent messages. Finally, we conclude by discussing some of the related work and proposing directions for future reearch.

2. EMAIL RESPONSE FOR THE INVESTOR RELATIONS DOMAIN

Our work on email response is based on a corpus of messages exchanged in the Investor relations domain, *i.e.* the process by which a company communicates with its investors. In this domain, electronic mail is used by enterprises in two different ways. First, outbound messages are sent by the company for the promotion of corporate events and financial results. Second most companies have, on their Web site, an investor relations section where financial analysts can be contacted to provide assistance to investors (inbound messages). Information available from these sites includes various topics such as stock prices and financial reports. These services are of significant importance as the quality of the information plays an important role in the decisions of professional investors. In the next paragraphs, we discuss some of the issues pertaining to this domain, our message corpus and its exploitation.

2.1 Some characteristics of the domain

Request messages are usually sent by corporate and individual investors as well as financial analysts. These messages are processed by a group of investor relations analysts. The requests pertain to various topics such as:

- *The characteristics of the corporation:* such as its ownership, its subsidiaries, the stock composition, etc.
- *Its financial results:* such as earnings, dividends and various financial ratios.
- *Its stock market performance:* fluctuations and stagnation, stock split, transactions/acquisitions, accounting practices.
- *Occurrence of events:* details on conference calls, dates of the publication of quarterly reports.

Other messages are also received about various aspects like taxation, debentures, updates of personal accounts, definitions of financial terminology and complaints about the web site.

2.2 Some characteristics of the messages

In order to conduct our research, Bell Canada Enterprises (BCE) provided us with a corpus of inbound messages (over 1500) comprising both the requests submitted by the investors and the replies formulated by the analysts. In the following paragraphs, we discuss some of the particularities of these messages and their impact on our selection of an adequate CBR approach.

Message size: the average length of a message is approximately 87 words (varying between a few words and 157). Since the requests are sent by different investors, the style varies from one request to another. On the other hand, the responses, provided by 5-10 analysts, are somehow uniform in their format and structures. The responses are well written using an adequate vocabulary. Very seldom do we encounter sentences comprising negation of propositions.

Structure of incoming requests: Usually, the incoming requests are composed of parts:

- a header containing the date, the sender's address and the subject;
- a short description of the context: For instance, reasons why the email was written, like "I am considering investing in your company" or "I am conducting an analysis of your stock"; very seldom is there a detailed description of a problem being confronted by the investor;
- one or more sub-requests pertaining to the topics described in the previous section on domain characteristics;
- the coordinates of the investor: name, title, affiliation, postal and electronic address, and signature.

Sequences of messages: Seldom did we encounter multiple exchanges between analysts and investors. Most of the individual requests contain

sufficient information for the analysts to formulate adequate responses. Hence, we assume that the messages can be considered independent.

Specificity of the messages: the degree of specificity of the questions and response varies greatly. Some requests are generic; for instance, "Why should I invest in your company?" At the other end of the spectrum, others are very specific such as "since 1999 earnings for BCE are $8.35 per share and Nortel's are -$0.23, I would assume that after the spin-off, BCE earnings should be something more than $8.50 per share...". The specificity of a message can be measured by the proportion of pronouns/determinants at the first person, numeric quantities and proper nouns [Kos01b]. As for the responses, they do not always address directly the requests of the investor. For instance, speculations about stock market fluctuations receive a standard courtesy reply.

Multiple requests: an important characteristic of our corpus is that some messages contain more than a single request. For instance, an investor might ask for the last earnings report and also request to be added to the company's distribution list. This multiple request feature is one of the main differences between email messages and frequently-asked questions (FAQ) which normally pertain to a single topic.

Temporality: the content of a message refers to specific time periods. For instance, financial quarters can be described explicitly ("the third") and implicitly ("next", "previous"). The date of a message is also necessary to determine the context of a request containing implicit time references.

2.3 Our CBR Approach to Email Response

In the first phase of the Mercure project, a combination of information extraction and text generation techniques were studied to build an initial email response system. The corpus used in this phase depicted email correspondence exchanged for printer troubleshooting. However, since the investor relations domain presents more diverse situations and discussions that are less factual in nature, the strict usage of information extraction for analyzing incoming messages would present some difficulties in the second phase of the project. It would be a considerable task to predict the various situations and their corresponding extraction templates. Also as each situation occurs a limited number of times, it might reveal impossible to elicit robust extraction rules to fill each of these templates. Similar limitations would also apply to the template-based generation of responses.

As part of the second phase of the project, we are developing a case-based reasoning module to synthesize "analyst-like" responses. Presented from a client perspective, the CBR module attempts to reuse messages in the SENT mailbox of the analyst's email software to suggest responses to new messages incoming in the INBOX. As illustrated on Figure 1, the "search and adapt" reasoning scheme is extended to take into account the following tasks: retrieval of messages with multiple requests, extraction and substitution of the named entities, and identification of relevant passages in the response. These aspects are discussed in sections 3 and 4 of this paper.

To illustrate our CBR approach to email response management, let's consider the following request pertaining to the dissemination of financial

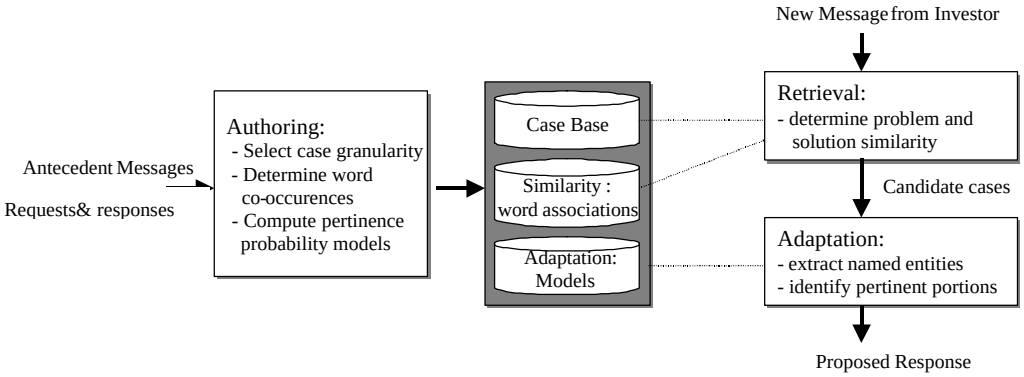

Figure 1 – Textual CBR processes for Email response

results (Figure 2):

Request$_1$: Can you tell me when you are reporting next. Thanks, Elliott.

Figure 2 – Example of an incoming request

The system must now generate a response to Request 1. Assume that the case j (request/response pair) from Figure 3 was selected by the retrieval phase. Given the textual characteristics of our application, we must next identify the relevant text passages of the Response$_j$ for reuse purposes.

Request$_j$: Hello
Can you tell me when you will be releasing your next earnings report also when your fiscal year ends
Best Regards, Mark Strasse

Response$_j$: Dear Mr. Strasse,
The year ended on 31 December 1999.
The release date for the next earnings report is on 26 January 2000.
Please, do not hesitate to contact us for any other questions. Sincerely…

Figure 3 – Case used for responding to Request$_1$[3]

Using message j, the CBR module determines that the passage pertaining to the earnings release date is useful. It then establishes that the passages related to the end of the fiscal year and the date of the quarterly report should be modified or simply pruned. The passages to be modified or pruned are indicated by the notation « text ». Hence, with substitutions recommended as tooltip text, the response proposed by the system would be the following (Figure 4):

This example illustrates the need to extend the CBR process beyond its retrieval phase in order to delay the intervention from the analyst to the latest possible stage of the response generation. Hence, the main challenges to our CBR module are the precision of the retrieval and the adaptation to improve response quality.

3. AUTHORING AND RETRIEVAL OF CASES

We present below two of the main issues to be considered as part of the case base authoring and the retrieval of pertinent cases.

3.1 Cases Granularity

A natural approach to the authoring of the case base is to associate a previous message (from the existing corpus), comprised of the request of an investor and the reply of an analyst, with a single case. However, retrieval of such antecedent messages is complicated by the fact that individual messages might pertain to various themes and contain multiple requests. This phenomenon seldom arises in case-based reasoning systems where the "single-fault assumption" is presumed; *i.e.* cases are normally comprised of a single problem description and its corresponding solution. However, in our application, this assumption does not hold if we are to associate messages to cases in a one-to-one fashion.

This brings us to the question of determining the granularity of a case, *i.e.* the mapping between a message's sub-requests and cases. We have considered three possibilities:

1:N Mapping: a case corresponds to a message, which may be comprised of multiple questions and responses. This offers the advantage that the messages are directly exploitable without major modification. However, the similarity of messages with multiple messages sub-requests is more difficult to determine. A diversity criterion [Smy01] could be used to cover the themes with a minimum number of cases.

1:1 Mapping: a case can only contain a single question and its associated response. This mapping offers the advantage that the similarity is established on portions of a text pertaining to a single theme. However, this would requires the usage of parsing techniques to segment the original messages. In our corpus, sub-requests are associated to sentences as follows:

Elliott

Response$_1$ Dear «Mr. Strasse»,

«The year ended on 31 december 1999».

24 October 2001

The releasedate for the next earningsreport is on «26 January 2000».

Please, do not hesitate to contact us for any other questions. Sincerely…

Figure 4: Response proposed by system.

[3] Names of individuals in the original messages were modified.

- the sub-requests are distributed over different sentences (the most frequent);
- a single sentence contains a conjunction of sub-requests; for instance the question presented in our previous example;
- a sentence contains a conjunction or an enumeration in a noun phrase; for instance, questions of the form "I would like to obtain the next reporting dates of BCE, Bell Emergis and Teleglobe."

Given a new question to which a response is to be generated and a fragmented case base, the retrieval phase would be conducted by successive searches over the various themes. We thus postpone the difficulty of producing a final result to a later stage where the individual retrieved responses must be combined together.

M:N Mapping: a group of disjoint cases are constructed from messages pertaining to common themes; (*i.e.* M cases covering N themes). The clustering of the themes, combined with a segmentation scheme as discussed for the 1:1 mapping, would result in a more compact case-based with less redundancy. However, it would be difficult to combine the individual portions to produce a meaningful response. For this reason, we did not adopt this direction in our current work.

3.2 Extending the scope of retrieval

During our initial experimentation, the similarity between messages was established based on the comparison of a tf*idf vectorial representation of the message content. We opted for a 1:N case mapping using the original messages. All messages are tokenized, tagged with parts-of-speech, and lemmatized (a morphological analysis of terms). Preliminary experiments indicated that such a scheme, using a cosine function for computing global similarity, provides a precision of approximately 57%, which resembles results from similar experiments with FAQs [Bur97].

However, the nature of our cases can be exploited to improve some aspects of the retrieval phase. As the selection of wrong answers requires additionnal manipulation by the user of the system, it is important to optimize the ranking of the most relevant(s) case(s) to ensure the production of a relevant response.

We considered two possibilities for improving the performance of the retrieval phase:

Using word relationships: similarity established on a vectorial representation of the cases has some limitations, as it requires the exact correspondence of words (or keyphrases or ngrams). To overcome this constraint, some authors [Bur97][Bru99] have made use of existing linguistic resources (*e.g.* thesaurus) to establish the semantic similarity of different words that have related meanings. This approach does not transpose well to our problem as, to our knowledge, no domain specific resources are available. More general linguistic resources, such as WordNet, do not provide a good coverage of the message terms. For instance, approximately 38% of the terms of our corpus are not covered (mostly financial terminology, proper names, companies …). However, as our case base is relatively substantial, we can obtain using experimental methods, an estimation of these relationships in the form of word co-occurrences. One advantage of using these co-occurrences, selected either through statistical tests or probabilistic measures, is their representativeness of the domain of discourse.

Exploiting textual responses: requests descriptions, written by different investors, present less uniformity than the responses provided by a limited number of analysts. Similarity should be more easily established when the textual responses are also taken into account during the retrieval phase.

We combined both of the above possibilities into a single scheme. A textual case can be seen as the linguistic "conversion" of a textual problem into a corresponding textual solution. The case base then corresponds to a mapping from a "request" language (problem) to a "response" language (solution). The finding of associations, captured as co-occurrences, provides indications that the occurrence of problem words increases the likelihood of the presence of some other words in the solution. To obtain the co-occurences, we collect the count of all pairs of words coming respectively from the requests and their corresponding responses, and we select the most significant ones based on the mutual information metric [Man99].

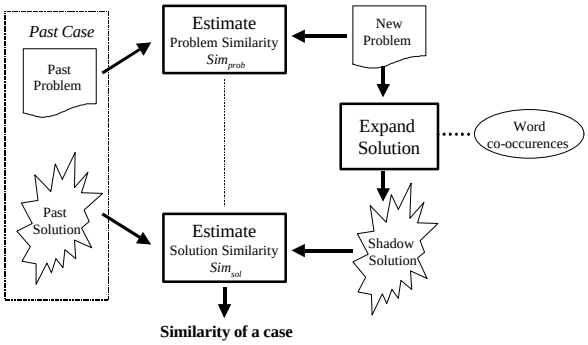

Figure 5: Exploitation of word-co-occurrences for solution expansion.

The approach we are currently using to insert the associations in the retrieval phase is inspired from

query expansion techniques. The incoming problem description (the investor's request) is expanded into a vector of response terms provided by the lists of co-occurrences. As illustrated on Figure 5, similarity of the cases then corresponds to the weighted sum of both problem and solution vector cosine.

Experimentation conducted on 102 test requests indicates that the expansion scheme slightly improves the overall precision[4] (62.0% vs. 57.9%) of the retrieval phase and preserves the rank of the first pertinent solution in the similarity list (2.01 vs. 1.96). The most significant improvement has been observed for the test messages where the response is not directly addressing the request (e.g. redirection to a generic web site address following the request of specific documents or financial information). For this category of message, the precision is almost doubled (80.1% vs. 51.0%) and the average rank is reduced to a very good level (1.33 vs. 2.38). For the other messages, the precision is mostly preserved but we observed some degradation for the routine messsages as the expansion scheme introduces some noise in the internal representation of the textual cases. This result is however interesting as responses are built from a limited number of the most highly ranked cases (usually the first one). And, most importantly, we expect that the selection of a judicious trade-off between request and solution similarities will bring further improvement.

4. REUSE OF TEXTUAL CASES

As mentioned previously, our application presents strong incentives to implement some adaptation of previous responses. While complete reformulation of past textual responses for diverse situations is beyond the capability of current CBR and NLP techniques, some of these techniques can nevertheless help to:

a) personalize past messages and
b) preserve the relevance of cases with the context of the new incoming request.

In the CBR literature, case adaptation (i.e. case reuse) has exclusively been conducted for structural cases and mostly corresponds to modifying the values of pre-selected solution features. In a textual setting like our email response domain, such a scheme is rather difficult to implement, as the textual solutions are not structured. Therefore, prior to the modification of the content of the messages, we need to determine what portions of the responses are promising candidates for modification.

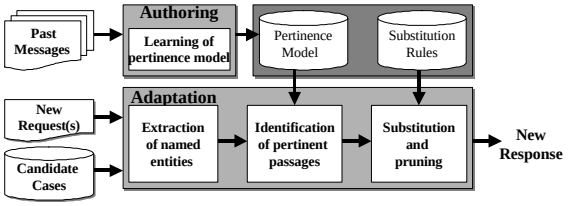

Figure 6 – Steps of the textual adaptation process

Given a new message and some past solutions selected during the retrieval phase, we implement the reuse of textual cases as a three-step process as illustrated on Figure 6:

- *Identification of relevant passages:* This corresponds to determining the text portions that are not applicable in the context of the new incoming request. Statistical distributions, captured as word alignments [Bro90], can be used for this task; further details are provided in section 4.1.
- *Message personalization:* among the relevant sentences, determine what text portions are subject to be modified. Usage of information extraction techniques [Cow96] for this step is discussed in section 4.2.
- *Pruning and substitution:* this corresponds to the removal of irrelevant passages and the substitutions of the portions to be personalized.

4.1 Identification of relevant passages

The presence of non-relevant passages is due to the occurrence of multiple themes in the requests and responses. The identification of these passages corresponds to the production of a subset of the precedent response based on the context of the new incoming request. In NLP, this corresponds to a query-relevant summarization process [Mit00], more specifically to the condensation of a text based on the terms of a request. As illustrated on Figure 7, the resulting solution R_c can be produced by removing, from the original response R, sentences (or text portions like noun phrases) that cannot be "matched" with the new incoming request Q.

[4] Precision is estimated as the percentage of pertinent responses contained in the first 5 nearest cases. The results presented for the expansion scheme are based only on the similarity of the solutions (responses).

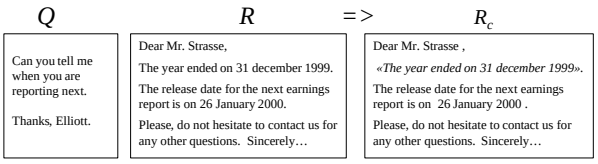

Figure 7 – Identification of relevant passages as a text condensation process

The matching process can be seen as a statistical process where we try to find the subset of *R* that best fits the request *Q*. In terms of probability, we are trying to find a condensed response *R'* that maximizes the following probability estimate

$$R_c = f(Q,R) = \arg\max_{R'} P(R'|R,Q),$$

Using bayes rule, this expression can be approximated as follows:

$$\begin{aligned} R_c &= \arg\max_{R'} P(R'|R,Q) \\ &= \arg\max_{R'} P(Q|R',R)P(R'|R) \\ &\sim \arg\max_{R'} P(Q|R')P(R'|R) \end{aligned}$$

Therefore, this step of the adaptation process corresponds to finding a compromise between a subset of the response that best "fits" the new request while preserving as much as possible the integrity of the past response.

The expression *P(R'|R)* can be modeled either as the coverage of both text or as a random drawing of words from the original request R. Some probability distributions (*e.g.* hypergeometric) can be used to evaluate the probability of the resulting condensed text. What is more critical is the granularity of text to be drawn from the original response. In our experimentation, we remove complete sentences as well as conjunctive forms of noun phrases. This strategy was selected to keep verb phrases intact and hence preserve the structure of the text.

The probability *P(Q|R')* corresponds to the probability that a request *Q* was at the origin of the response *R'*. A corpus is needed to learn such probability distributions. We are currently exploring some techniques to produce synthetically such a corpus form our case base. It would then be used to estimate the probability distribution *P(Q|R')* by applying some learning techniques (based on EM algorithm) as described in [Bro90]. This approach offers the advantage of being domain independent and transposable to other domains where a sufficiently large corpus is available.

4.2 Personalization of the messages

Based on observations made from our corpus, personalization of messages refers to the capacity to detect some factual information in the messages and to substitute them. This includes for instance, the names of companies, individuals, financial factors, dates and time references. These correspond to named entities and can be identified using information extraction techniques (IE). IE techniques identify, using either rule patterns or statistical models, information from textual documents to be converted into a template-based representation. As during the first phase of the project, we make use of extraction patterns and lexicons (lists of company names, titles, acronyms and frequent financial terms).

Substitutions of these entities are partly conducted using a rule-based approach. Replacement of individual names and companies is based on the roles of the messages entities. The role is determined by the type of patterns used during extraction, mostly based on the part of-speech and the terms preceding/following the entities. For instance, expressions like "*Sincerely, John Smith*", "*to purchase Nortel shares*", "*registered with Montreal Trust*", could provide indications of the message sender, subsidary company and financial institution respectively. However, as the Investor Relations domain does not offer much predictability, the elicitation of domain rules for numeric information (dates, price, factors…) remains difficult and such substitutions rely mostly on the user.

5. RELATED WORK

Current email management technology provides capacities for message categorization, routing (assignment to the right individual or department), queuing (priority establishment) as well as some static response capabilities. While few companies provide detailed information about their underlying technology, it appears that most are based on classification techniques such as neural networks. Conversational CBR has been adopted by a few systems to provide pre-determined static responses (*e.g.* former Inference K-Commerce Email product).

Shimazu and Kusui [Shi01] have proposed an approach for the detection of new cases based on the increase (and decrease) of keywords from messages in a call tracking database. Their system, SignFinder, helps in the decision of creating "frequently-asked questions" type of cases; however it does not address the formulation of textual responses. These authors also proposed a technique for finding relations among messages on a bulletin board [Kus01]. This differs from our work since sequences of messages seldom arise in our

application domain, which lead us to assume the independence of the Investor Relations messages.

Our choice of a textual CBR model as opposed to a conversational CBR model brought up some interesting issues. The conversational model [Aha01] relies on the hypothesis that the user must provide answers to questions asked by the system. This implies that, in our application, either the investor or the analyst would provide these answers. If the investor does provide answers, it should be accomplished through the company's web site and then would bring some undesired change to the current mode of interaction with the analysts. If the analyst does provide answers, the interaction with the conversational system might be more demanding than writing directly a response (as they contain on average ~30 words). Also as most of the request descriptions are explicit, it would seem odd to ask the user to guide the retrieval phase with redundant questions. For these reasons, we have deemed that a textual model is most suitable for our email response task. As the questions submitted to the analysts are sufficiently clear and explicit, we believe that such a process can be efficiently sustained by a textual CBR model augmented with adequate NLP techniques.

6. CONCLUSIONS

In this paper, we have presented our approach for the development of an email response module based on CBR techniques. Our study revealed some requirements pertaining to the personalization of messages, the modification of messages containing multiple themes and the needs for precision in the retrieval phase. We have discussed three issues to be addressed within the CBR cycle, namely:

a) the choice of granularity for cases and its impact on the retrieval phase;
b) the use of textual responses in the retrieval phase;
c) the modification of previous responses based on a statistical model and information extraction techniques.

As for future work, we are experimenting with another statistical approach (translation model) for the retrieval phase. We expect these models to be more selective and to reduce the level of noise generated during the expansion process. One key issue will be to determine how sparse data affect the quality of the relevance identification process that we proposed.

Acknowledgements. The completion of this research was made possible thanks to Bell Canada's support through its Bell University Laboratories R & D program.

REFERENCES

Aha, D., Breslow, L., Muñoz-Avila, H., 2001. Conversational case-based reasoning. Applied Intelligence, 14, pp. 9-32.

Branting, L., Lester, J., 1996. Justification Structures for Document Reuse, In *Proceedings of EWCBR-96*, Lecture Notes in Artificial Intelligence 1168, pp.76-90.

Brown, P., Cocke, J., Della Pietra, S., Della Pietra, V., Jelinek, F., Mercer, R., Roossin, P., 1990. A Statistical Approach to Machine Translation, Computational Linguistics, 16(2), pp. 79-85.

Brüninghaus, S., Ashley, K., 1999. Bootstrapping Case Base Development with Annotated Case Summaries, In *Proceedings of ICCBR-99*, Lecture Notes in Computer Science 1650, Springer Verlag, pp. 59-73.

Burke, R., Hammond, K., Kulyukin, V., Lytinen, S., Tomuro, N. & Schoenberg, S., 1997. Question Answering from Frequently-Asked Question Files: Experiences with the FAQ Finder System. AI Magazine, 18(2), pp. 57-66.

Mittal, V., Berger, A., 2000. Query-relevant summarization using FAQs,. In Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL). Hong Kong.

Cowie, J., Lehnert, W., 1996. "Information Extraction," Communications of the ACM, vol. 39 (1), pp. 80-91

Gartner Research, 2000. E-mail Response Management: Perspective, http://cnscenter.future.co.kr/resource/rsc-center/gartner/email.pdf.

Kosseim, L., Beauregard, S., Lapalme, G., 2001, Using Information Extraction and Natural Language Generation to Answer E-mail, in Natural Language Processing and Information Systems, Lecture Notes in Computer Science 1959. Springer-Verlag, pp. 152-163.

Kosseim, L., Lapalme. G., 2001. Critères de sélection d'une approche pour le suivi automatique du courriel, In *Actes de la 8ème conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN-2001)*, Tours, France, pp. 357-371.

Kusui, D., Shimazu, H., 2001. Transforming Mailing Lists into Case Bases, In *Proceedings of ICCBR'01*, Lecture Notes in Artificial Intelligence 2080, pp. 690-701.

Manning, C., Schütze, H., 1999. *Foundations of Statistical Natural Language Processing*, The MIT Press, Cambridge, Massachusetts.

Shimazu, H., Kusui, D.; 2001. Detecting Defect Sign Cases, In *Proceedings of ICCBR'01*, Lecture Notes in Artificial Intelligence 2080, pp. 611-621.

Smyth, B., McClave, P., 2001. "Similarity vs. Diversity", In *Proceedings of ICCBR'01*, Lecture Notes in Artificial Intelligence 2080, pp. 347-361.