Using Statistical Word Associations for the Retrieval of Strongly-Textual Cases

Luc Lamontagne, Philippe Langlais and Guy Lapalme

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

Abstract

Lexical relationships allow a textual CBR system to establish case similarity beyond the exact correspondence of words. In this paper, we explore statistical models to insert associations between problems and solutions in the retrieval process. We study two types of models: word co-occurrences and translation alignments. These approaches offer the potential to capture relationships arising between a problem description and its corresponding textual solution. We present some experimental results and evaluate these with respect to a tf*idf approach.

Introduction

Recent work, grouped under the "Textual CBR" banner, has devoted much interest to the exploitation of cases described by textual documents. However, most of these studies concentrate on the textual descriptions of problems while very little attention is paid to the solutions descriptions. Our objective is to evaluate the contribution to the CBR reasoning cycle of the relationships between problems and solutions, when both are textual. We refer to these cases as being "strongly-textual". This is particularly of interest to applications such as the answering of frequently-asked questions (Burke et al. 1997), the response to email messages (Lamontagne and Lapalme 2003) or the management of diagnostic reports (Varma 2001). In these applications, both the content and the narrative form of the solution are of interest to the problem solving approach.

The motivation behind our research is to determine whether the modeling of the lexical relationships between problem and solution components may be inserted into some of the phases of the textual CBR cycle in order to improve overall system performance. In this work, we concentrate our efforts on the retrieval phase. The CBR literature on the retrieval of textual cases has mainly been inspired by techniques studied in information retrieval systems. Most of these efforts make use of a vectorial representation of the cases comprising keywords (Burke et al. 1997), character ngrams (Aha 2001) and keyphrases (Wilson 00). Similarity between a problem description and candidate cases is normally established using a cosine product of the term vectors. However, this approach has some limitations as it requires the exact correspondence between terms (or character ngrams). To overcome this constraint, some authors (Burke et al. 1997) (Brüninghaus and Ashley 1999) proposed the use of general-purpose linguistic resources (*e.g.* thesaurus) to establish the semantic similarity of different words with related meanings. While it provides some improvements, this approach might also pose some problems since the notion of semantic similarity is rather difficult to establish and since linguistic resources, such as WordNet, are often quite too general for the domain being addressed.

In this paper, we investigate two approaches based on statistical natural language processing (NLP) techniques to determine what methods provide the most promising results for improving retrieval. Our goal is two-fold: To evaluate the benefits of aligning problem and solution descriptions in the retrieval process, and to compare this approach with those more commonly used for retrieval in textual CBR. To conduct our study, we make use of some resources available from our current application, email response for the investor relations domain.

The main idea underlying our approach is that a textual case represents the lexical "conversion" of a problem description into a corresponding solution description. The case base then forms a corpus of parallel texts (a bitext) and statistical methods allow for the finding of associations, captured as statistical models, among words from both problem and solution descriptions. We study two types of statistical models: word co-occurrences and translation alignments. Word co-occurrences provide some indications that the occurrence of problem words increases the likelihood of the presence of some other words in the solution. Statistical alignments impose stronger relationships as each word in a problem is assumed to be the direct translation of a single solution word.

In the following sections, we describe the two statistical models and the corpus used for our experiments. We then present some results and finally we propose some directions for future work on this research theme.

Making use of Problem–Solution Associations

The ranking of textual cases based on their relevance to a new target problem is normally estimated by the similarity of problem descriptions. For some application domains, many elements of the relationship between problems and solutions may further be exploited in the CBR cycle. For instance, let us consider some of the following issues:

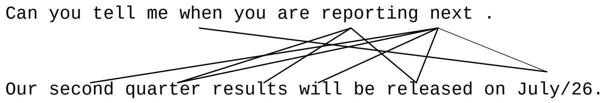
- The uniformity in the writing of the solutions is greater than that of the problem descriptions. For example, in the investor relations domain, solutions are written by a limited number of financial analysts as opposed to problems which have been submitted by different investors (corporate and individual) with different background and experience of the financial market. Usage of solutions may then provide a more homogeneous way of comparing cases.
- The formulation of a solution mimics the description of a problem. In this case, the textual solution is composed to address the various portions of the situation description. Consequently, a correspondence may be established between paragraphs, sentences or syntactic phrases of both case components. Once again, we can expect benefits from exploiting these mappings.
- It may be preferable to select cases with textual solutions easier to reuse. For instance, one might prefer a compromise between problem similarity and solution length, since shorter ones are easier to modify.

In order to incorporate case solutions in the retrieval phase, a possible naïve approach is to merge words from both problems and solutions in one case internal representation and to use this combined structure to estimate case similarity. However, some limitations can be anticipated with this approach. First, there is no guarantee that the same vocabulary is used to describe both problems and solutions. These might have been written by different individuals with different backgrounds and might represent different domain perspectives. For instance, a novice investor seeking help would hardly refer to financial indicators that would be used by professional analysts. Also, a situation description might not be directly addressed in the solution. For example, the response to some financial requests might be to consult a web site or read some documents. Therefore, in order to exploit such case characteristics, we need tools to represent the lexical transfer from the problem descriptions to the solution components.

In our framework, we assume that the textual case base is heterogeneous, *i.e.* that it depicts various situations and proposes diverse solutions dissimilar in nature. Our proposal to use lexical associations to discriminate among cases might have a smaller impact for homogeneous case bases where a large portion of the words is repeated in most cases.

Exploiting Word Co-occurrences

The first approach we use to insert word associations in the retrieval phase is inspired from query expansion techniques. We make use of co-occurrences, which indicate that the presence of specific words in problems should increase the likelihood of finding some other specific words in solutions. Hence a given problem word can influence the occurrence of several solution words, the converse being also possible. For instance, in the following pair,



```
Can you tell me when you are reporting next .

Our second quarter results will be released on July/26.
```

co-occurrences capture, with different association strengths, that "reporting" is related to "results" being "released" and that "July/26" is "when" the "next" report will come out.

To obtain the co-occurrences, we counts the frequencies of all pairs of words (w_{prb}, w_{sol}) that come respectively from the problems and their corresponding case solutions, and selects the most significant ones based on the mutual information measure (Manning and Schütze 1999). Mutual information, expressed as in the following equation, indicates the amount of information one word provides to the other:

$$I(w_{prb}, w_{sol}) = \log \frac{P(w_{prb}, w_{sol})}{P(w_{prb})P(w_{sol})}$$

For each problem word w_{prb}, the various w_{sol} are then ranked by decreasing order to form co-occurrence lists. These lists are truncated using thresholds based on the mutual information value and on the number of words associated with the same term.

To make use of these co-occurrences in the retrieval phase, the target problem is converted into a new structure using terms provided by the co-occurrence lists. This "expansion" function provides a vectorial representation of a solution approximation, which we refer to as the "shadow" solution. The shadow solution corresponds to a weighted average of the co-occurrence lists for each word present in the target problem. To each term in a list, we associate the weight of its corresponding word w_{prb}. Finally, the similarity of the shadow vector with the case solution is estimated using the cosine function.

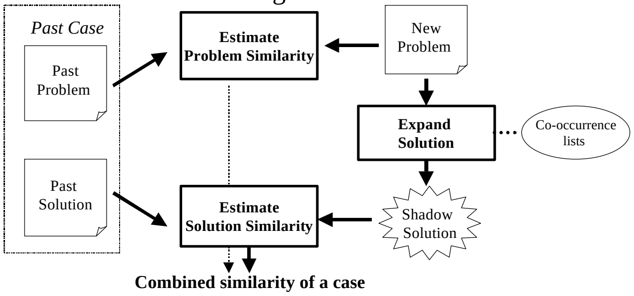
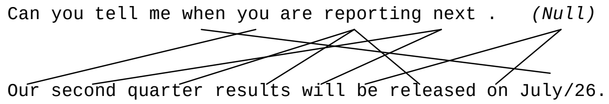


Figure 1 - Case similarity using co-occurrence lists

Similarity Estimates with Translation Models

The second approach we study, namely alignment models, is used in statistical machine translation. The idea of aligning two sentences consists of determining, for each word in one sentence, the words of the other sentence from which it originates. In our previous example, a translation approach would impose that each word of the second sentence (the target) be generated from a single word from the first sentence (the source). Such models also admit that target words can be generated from nothing (the *null* word) for grammatical purposes.


Can you tell me when you are reporting next . (Null)
Our second quarter results will be released on July/26.

Transposing this idea to textual CBR, one can imagine that there exists a language for describing problems and another for describing solutions. Hence, a case can be viewed as a translation of a situation description into some solution language. The models governing this translation, learned from our case base, could then be used to rank the pertinence of a previous solution with respect to a new problem. However, the generation of a new solution description based on these models is not envisaged, as it would be stretching the analogy far too much.

Various models are proposed in the NLP literature for computing such models. The IBM models (Brown et al. 1993) were developed for learning from a corpus the probabilities necessary to establish the alignments between parallel texts. These models are of increasing complexity and take into account various factors such as the generation of multiple words, the distortion of word positions and the grouping of words. For this experiment, we made use of the model IBM1 that can be formulated as follows in our CBR framework: The probability of finding a problem *Prob* given a solution *Sol* is:

$$p(Prob \mid Sol) = \sum_{a} p(Prob \mid a, \mathrm{Sol})\, p(a \mid Sol)$$

This corresponds to the probability of obtaining a sequence *Prob* through all the possible alignments α with the sequence *Sol*. Following some manipulations and simplifications, the conditional probability for this model is expressed as:

$$p(Prob \mid Sol) = \frac{e}{(l+1)^m} \prod_{j=1}^{m} \sum_{i=1}^{l} t\left(Prob_j \mid Sol_i\right)$$

where the sequences *Prob* and *Sol* contain respectively *m* and *l* words. This expression is used both for the ranking of the solutions and for the learning of the translation model. The result of the learning process is the transfer table *t* that provides probabilities of generating a target word $Prob_j$ given that a word Sol_i is present in the source description. The transfer model can be found by applying an EM algorithm that iteratively assigns and revises probabilities to the model parameters until convergence is reached.

It might seem odd to the reader that we consider problems to be generated from solutions and not the converse. Apart from some aspects related to the noisy channel model underlying this approach, the main reason is that probabilities are multiplicative in nature and comparing solutions of different lengths would favor those with fewer terms. By ranking solutions for their contribution to a fixed length problem description, we ensure that they are evaluated on a common ground.

Our Corpus

The work presented in this paper is part of a project on email response using CBR techniques. We are currently using a corpus of messages[1] pertaining to the Investor Relations domain, *i.e.* the process by which a company communicates with its investors. The corpus consists of over a thousand inbound messages exchanged between investors and financial analysts. The messages cover a variety of topics such as requests for documents, financial results, stock market behavior and corporate events.

A case consists of a pair of messages, a request from the investor and a response from the analyst. The length of the individual messages varies from a few to over 200 words with an average of 87 words. The responses, provided by few analysts (5-10), are somehow more uniform in their formats and structures than the requests sent by different investors. Most messages are well written using an adequate vocabulary.

In this study, we wished to investigate the capability of selecting cases based solely on the textual content of the messages. We therefore removed the headers (*e.g.* date, subject and recipient fields) and the non-textual MIME body parts. The words of the texts were tokenized, tagged with a part of speech and lemmatized to obtain their morphological root. Finally, we replaced dates, phone numbers, URL and email addresses by a semantic tag (*e.g.* DATE) to reduce the specificity of the messages and to favor a comparison of the messages on a common ground. As for the internal representation of the cases, we kept both a vectorial representation and the sequence of lemmatized terms that form the problem and solution components. To reduce the vocabulary of the case base, terms were filtered based on their corpus frequency and on the usefulness of their lexical category.

Experiments

We present the results of a comparison of the co-occurrence model, the translation model and the tf*idf similarity encountered in textual CBR. The results were obtained using a leave-one-out evaluation, for which a case is held out of the case base prior to training the models and used as a target problem for evaluating the retrieval phase. We used a subset of 102 pairs of messages grouped under the topic “financial information”. To gain

[1] The corpus was provided by Bell Canada Enterprise.

more insight into the nature of the results, we subdivided this test corpus into four topical groups:

A - Reporting of financial results and conference calls: Messages are uniform and use a limited vocabulary.

B - Requests about financial aspects: Messages are diverse and might contain detailed requests and explanations varying from one message to another. Some responses to speculative messages are generic and address them indirectly (*e.g.* "consult our web site").

C - Distribution lists: Requests from investors to join some distribution lists. However the adherence to the list is not always confirmed in the response.

D - Other messages: These cover various dissimilar topics, with few antecedents that could be used as a response.

To compare the approaches, we used three criteria:

- Average rank: The position of the first pertinent case in the list of nearest neighbors;
- First position: The proportion of trials for which the nearest neighbor is pertinent;
- Precision: The proportion of pertinent cases in the first k nearest-neighbors (k=5).

*tf*idf* Similarity

A common approach for measuring textual similarity is to compare the vectorial representations of a new problem *Prb* and the problem description of a case *C*. To each term of the vectors is assigned a weight *w(t)* determined from its frequency in the description (*tf*) and its relative distribution in the case base (*idf*) expressed as:

$$idf(t) = \log\left(\frac{|CB|}{|CB:\ \text{where}\ tf(t) > 0|}\right)$$

where |*CB*| represents the size of the case base.

Local similarity is restricted to identical terms and the global similarity is then established by the normalized cosine of both vectors:

$$sim(Prb, C) = \frac{\sum_t w_{Prb}(t) \times w_C(t)}{\sqrt{\sum_t w_{Prb}(t)^2 \times \sum_t w_C(t)^2}}$$

We obtained the following results on our corpus:

Group	Average rank	First	Precision
All	1.960	57.4%	57.1%
A	1.080	92.0%	80.0%
B	2.385	51.7%	51.0%
C	1.550	77.7%	62.2%
D	3.000	33.3%	30.3%

These results indicate that the overall precision is approximately 57%, and the nearest case is pertinent also for 57% of the trials. A pertinent case comes out first most of the time for messages of groups A and C. These case descriptions are routine messages with a limited vocabulary. However, lower performance is observed for groups B and D, which cover a wider range of topics.

To evaluate the potential benefits of using word associations in the retrieval process, we repeated this experiment using solutions instead of problems. An overall precision of 74.9% was reached. The nearest neighbor was almost always pertinent for groups A and C, and the average rank was significantly reduced for groups B (1.429) and D (2.083). This provides incentives to incorporate solution extrapolations in case retrieval.

Word Co-occurrences

The results obtained by generating shadow solutions with co-occurrence models are presented in the next table.

Group	Average rank	First	Precision
All	2.016	62.3%	62.0%
A	1.640	68.0%	69.8%
B	1.333	83.3%	80.0%
C	1.833	83.3%	66.7%
D	4.000	25.0%	35.0%

While the overall precision is superior and the pertinence of the nearest neighbor is improved for groups B and C, the performance deteriorates significantly for group A. We can explain these results with three observations. First, the co-occurrence model can detect when a solution is common to two different problems (*e.g.* a generic message used as a response to different speculative requests). This behavior explains some of the improvement for group B. Second, cases with longer problem descriptions tend to get better similarity values. More terms in the descriptions lead to more elaborate shadow solutions and hence cover more situations. Even after normalization, these solutions tend to get a higher ranking. In our test case base, the longer descriptions were weakly related to the other cases. Therefore, their presence in many of the nearest neighbor lists partly explains the degradation for group A. Our third observation is that while the co-occurrence model introduces valuable words in the shadow solutions, it also introduces noise. Examples of co-occurrence lists pertaining to group A (the "release of earnings report") are presented in the next table. The lists for "release" and "report" contains associations that are representative of the discussions in this group of messages (*e.g.* "schedule", "conference", temporal references). However, they also introduce less pertinent terms like "far", "also" and "detail". The result obtained for "Earnings" also reveals some limitations. This term is widely spread in the case base and is associated to many different co-occurrence pairs. The resulting list contains some associations that could contribute to the ranking of cases of group B but none to group A.

Prbl. word	Co-occurrence list of solution words
Release	earn, BCE_Emergis, CGI, EMAIL_ADDRESS, next, meeting, schedule, release, TIME, conference …
Earnings	EPS, reflect, study, read, such, accounting, note, analysis, next, holding, prior, item, give, also …
Report	next, day, give, quarter, far, TIME, DATE, detail, release, afternoon, after, reply, also, date, corporation, number …

Translation Model

For this experiment, we obtained an IBM1 transfer table using the GIZA++ toolkit (Och and Ney 2000) on our test case base. Using this model, the case ranking procedure provided the following results:

Group	Average rank	First	Precision
All	1.721	63.9%	56.9%
A	1.320	76.0%	74.4%
B	1.464	75.0%	58.6%
C	2.000	66.7%	60.0%
D	4.000	33.3%	25.0%

We observed a significant improvement in the average ranking of the first pertinent nearest neighbor. This is mostly explained by the results obtained for groups A and B. Coming back to our previous example, we observe that the new lists obtained with this model are limited to a few words, most of them being very pertinent. This model introduces less noise than the co-occurrence approach.

Probl. word	Translated from these solution words
Release	release, call, that, conference
Earnings	result, date, earnings, NULL, conference,next
Report	do, quarter, date, give

To explain the quality of the results of the first two groups, it is interesting to note the precision of some of the associations obtained through the translation model.

Problem word	Translated from these solution words
Distribution	list
Dial	PHONE_NUMBER, participate
Conference	conference, release, usually, dial

It is expected that further improvement will be reached as we extend the training of the model to a larger test case base. Nonetheless, this model already represents a good compromise between the results obtained by tf*idf similarity and the adequate usage of solutions.

Discussion and Future Work

In this paper, we proposed two statistical models for the retrieval of textual cases and compared their performance with a tf*idf similarity approach. Our experiments indicate that the translation model provides significant improvements in terms of precision and overall ranking. We observed that the co-occurrence model is an interesting approach in spite of the lexical noise it generates. These results illustrate the benefits of inserting word associations techniques into textual CBR systems.

The insertion of word relationships in the textual CBR cycle was also studied from a thesaurus-based perspective by (Burke et al. 1997) and (Brüninghaus and Ashley 1999). Our results indicate that the synonymy and hypernymy relations provided by a thesaurus are seldom encountered in the association lists and may be of very little use in depicting the relationships between problem and solution descriptions.

In the Information Retrieval (IR) community, statistical models are beginning to appear (Croft, Callan and Lafferty 2001) as language models, hidden-markov models and translation models are being investigated. However, IR tasks differ from CBR tasks as they do not make a clear distinction between problems and solutions. Such a partitioning provides opportunities for richer analysis in the application of these statistical models.

There are several directions in which this work can be extended. Instead of cumulating all the co-occurrences from problem-solution pairs, restricting the counting process within a limited-size window could reduce some of the noise. More extensive comparisons will be conducted to get a clear estimate of the performance improvement that can be reached for various case base characteristics. The size of the case base is of primary importance and we expect more representative lists to be generated from a larger corpus. It would also be of interest to conduct thorough experiments on texts of different sizes, domain dependence and levels of structuring.

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

References

Brown, P. F.; Della Pietra, S. A.; Della Pietra, V. J.; and Mercer, R. L. 1993. The mathematics of statistical machine translation: Parameter estimation, *Computational Linguistics*, 19(2): 263-311.

Brüninghaus, S. and Ashley, K. D. 1999. Bootstrapping Case Base Development with Annotated Case Summaries, in *Proceedings of ICCBR-99*, Heidelberg, Germany: Springer Verlag.

Burke R.; Hammond K.; Kulyukin V.; Lytinen S.; Tomuro N.; and Schoenberg S., 1997. Question Answering from Frequently-Asked Question Files: Experiences with the FAQ Finder System, Technical Report TR-97-05, Dept. of Computer Science, University of Chicago.

Croft, B.; Callan, J.; and Lafferty, J. eds. 2001. *Proceedings of Workshop on Language Modeling and Information Retrieval*, Carnegie Mellon University, http://la.lti.cs.cmu.edu/callan/Workshops/lmir01/

Lamontagne, L. and Lapalme, G. 2003, “Applying Case-Based Reasoning to Email Response”, Forthcoming.

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

Och, F. J. and Ney, H. 2000. "Improved Statistical Alignment Models". Proc. of the 38th Annual Meeting of the Association for Computational Linguistics, 440-447.

Wilson, D. C. and Bradshaw, S. 2000. CBR Textuality, *Expert Update*, 3(1):28-37.

Varma, A. 2001. Managing Diagnostic Knowledge in Text Cases” In Aha, D., and Watson, I., eds., *Proceedings of ICCBR’2001*, LNAI 2080, 622–633, Berlin: Springer.