Using Statistical Translation Models for Textual CBR

Luc Lamontagne

Department of Computer Science and Software Engineering
Laval University, Québec, Canada, G1K 7P4
luc.lamontagne@ift.ulaval.ca

Abstract. Textual CBR has developed as an important area of case-based reasoning as cases for many applications can only be made available in textual format. One of the challenges for reasoning with textual cases is the efficient exploitation of solution descriptions. Due to the lack of structuring of solutions expressed in natural language, it often reveals difficult to identify relevant cases and to conduct solution adaptation. In this paper, we examine how techniques from the field of statistical machine translation could contribute to some of these problems.

Keywords: Textual CBR, Statistical machine translation, Alignment models, Case retrieval, Textual adaptation.

1 Introduction

Textual CBR is a subfield of case-based reasoning dedicated to the exploitation of cases described by textual documents. Textual cases, expressed in natural language, come in varying forms and often exhibit characteristics which make the implementation of reasoning schemes challenging. One of these issues is the importance of solutions descriptions in the textual cases. Due to their lack of structuring, it reveals difficult to exploit the content and the narrative form of the solutions in the problem solving approach.

Our interest is to study how the relationships between problems and solutions, when both are textual, can contribution to the CBR reasoning cycle. More specifically, the modeling of lexical relationships between problem and solution components may be inserted into some of the phases of the textual CBR cycle in order to accomplish some tasks and improve system performance.

Statistical machine translation (SMT) is a specialized area of natural language processing (NLP) that provides techniques for modeling relationships between pairs of texts expressed in different languages. Such a framework can be transposed in textual CBR. The main idea underlying this approach is that a textual case represents the lexical "translation" of a problem description into a corresponding solution description. The case base then forms a corpus of parallel texts (a bitext) and statistical translation allow for the finding of associations, captured as statistical models, among words from both problem and solution descriptions. Statistical

alignments impose strong relationships as each word in a problem is assumed to be the direct translation of a solution word (or the converse).

The next section provides a description of how the relationship between problems and solutions components of a textual case can be exploited. The third section reviews the main elements of statistical translation techniques and explains how text alignment can be obtained. The formulation of textual CBR retrieval as a statistical alignment problem is discussed in the fourth section. Then Section 5 highlights how extensions of SMT can be applied to the adaptation of textual cases, and the paper finishes with some discussions in Section 6.

2 Relationship between Problems and Solutions

For some application domains, the relationships between problems and solutions may be exploited in the CBR cycle to enhance reasoning capabilities. For instance, let us consider some of the following issues:

- The vocabulary used to describe problems and solutions might differ significantly. In order to incorporate case solutions in the CBR cycle, 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. Also, a specific word might have a different meaning when used either as a situation descriptor or as a solution element.
- The uniformity in the writing of the solutions can be greater than that of the problem descriptions. For example, in customer relationship domains, solutions might be written by a limited number of employees as opposed to problems submitted by different customers with varying background and experience. Usage of solutions may then provide a more homogeneous way of comparing cases.
- The formulation of a solution is structured according to 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 that exploiting these mappings would facilitate case adaptation and reuse.
- It might be preferable to select cases with textual solutions that are easier to reuse. For instance, one user might seek a compromise between problem similarity and solution length, since shorter ones are easier to modify and reuse.

The benefits expected from modeling the relationship between problem and solutions varies with the properties of the cases. For instance, some textual cases contain structured solutions such as identifiers, classes or numerical evaluations. Spam filtering applications [1] are representative of this category. In such a context,

information retrieval and text mining techniques are usually sufficient for finding relevant cases and for recommending structured solutions.

On the other hand, cases containing short textual solutions can strongly benefit from these relationships. In these cases, we suppose that problems and solutions are identified as distinct parts and contain few sentences. Frequently asked questions [2], email exchanges [3] and some incident reports [4] have such a structure. Due to the limited complexity of their textual descriptions, retrieval and reuse of this category of cases can be sustained by word relationships. This is the issue we discuss in the rest of this paper.

Finally, we assume that such techniques could also be advantageous for cases originating from complex documents with long and ill-defined solutions. But problem and solution descriptions are often interleaved in the text, thus making the structuring of the cases even more complicated. Some efficient schemes would be required to disentangle sentences and associate them to their proper case component.

In order to exploit such case relationships, we need tools to represent the lexical transfer from the problem descriptions to the solution components. Statistical machine translation offers some techniques to capture and exploit word associations as statistical models.

3 Statistical Translation Models

Statistical machine translation [5] makes use of probabilistic models to generate translations from a source language to a target language. This paradigm, which gained significant popularity in the early 1990's, is now widely used (see for instance *Google Translate*). It offers some tools for tackling the following problems:

a) *Sentence alignment*: to establish the correspondence between sentences of two texts known to be a mutual translation. Sentence alignment algorithms use dynamic programming methods where a source sentence can be associated to 0, 1 or many other target sentences. Features such as sentence length and lexical information are exploited in this phase to determine the optimal pairing of sentences [6].

b) *Word alignment*: to find correspondences *a* between words in parallel sentences (sentences being aligned together). The task of a word alignment system is to indicate which word *f* in the source sentence generated the word *e* of the target sentence.

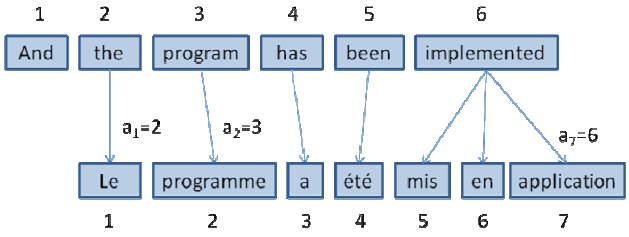

Fig. 1. An alignment between English and French sentences (adapted from [7]).

This phase rely on the acquisition of a probabilistic model (similar to a bilingual dictionary) indicating the possible translations for each individual word f and their respective likelihood $p(e|f)$. Unsupervised training is conducted on sentence aligned parallel corpora and is implemented as an expectation-maximization algorithm consisting of the following steps:

- Initialize the model $p(e|f)$ with uniform probability distributions;
- Apply this model to the training data and evaluate the most likely alignment by counting the number of times we find e in the target sentence given the presence of f in the source sentence (expectation step);
- Update the model from the data by re-estimating the probability distributions (maximization step);
- Iterate on the last two steps until some stopping criteria is satisfied (normally the convergence of the probability values).

The specifics for each of these steps depend on the model used to represent the transfer distribution. This is addressed further in this section.

c) *Sentence decoding* : to determine, based on a probabilistic model, which sentence e of a target language is the best translation of sentence f in a source language, i.e. to find

$$\tilde{e} = \arg\max_{e} p(e \mid f) = \arg\max_{e} p(f \mid e)\, p(e) \qquad \textbf{(1)}$$

where $p(e|f)$ is the translation model obtained from word alignment indicating the fidelity of the translation and $p(e)$ is a language model estimating the fluency of the sentence e. To avoid an exhaustive search by going through all possible sentence e of the target language, heuristic search is required to efficiently explore the translation search space.

In this paper, we restrict our treatment of SMT to the word alignment problem using word-based models. However it is important to note that recent progresses were made with phrase-based models [8], i.e. models where sequences of words are aligned together.

A series of models, referred to as the IBM models, were proposed in [5]. The Model 1, the simplest of the series, is characterized by the following assumptions:

- A target word can only be generated by a single source word.
- A target word can be generated without the support of any source word (represented as the NULL word).
- An alignment only depends on the words and does not take into account the position of the words in their respective sentences.

As these assumptions are normally too restrictive for modeling translation distributions, four other models of increasing complexity were proposed to take into account the length of the sentences and the position of the words being aligned (Model 2), the fertility of a word linked to many others (Model 3), and the permutation and the grouping of words (Models 4 and 5).

In the next two sections, we discuss how statistical translation models can be used for the retrieval and adaptation of textual cases.

4 Retrieval in Textual CBR using Translation Models

The CBR literature on the retrieval of textual cases has been strongly inspired by techniques used in information retrieval systems. Most of these efforts make use of a vectorial representation of the cases comprising keywords, character ngrams and/or keyphrases. 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.

Statistical machine translation, and namely alignment models, can help to overcome this constraint. 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. Transposing this idea to textual CBR, one can imagine that there exists a language for describing problems and another for describing solutions (Fig. 2). 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.

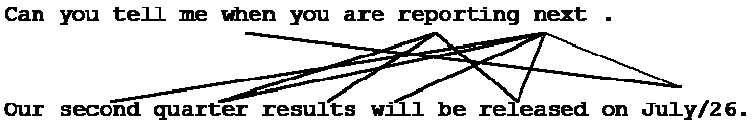

Fig. 2. An example of a word alignment between problem and solution descriptions.

The IBM Model 1 [5] presented in the previous section is adequate for acquiring such a model. Experimental results presented in [9] clearly revealed that, while being more conservative than approaches based on co-occurrence, it allows for the capture of meaningful word associations even with a restricted number of cases. One could think that this model is too simple and that linguistic phenomena, such as fertility and word permutation, are important. But, as the alignment of problem and solution descriptions tends to be sparse, these do not seem to be a determining factor. And the faithfulness of the resulting alignment is less crucial for textual CBR applications than for machine translation systems.

The IBM Model 1 can be formulated as follows in our CBR framework. The probability of finding a problem *Prob* given a solution *Sol* is:

$$p(Prob \mid Sol) = \sum_{a} p(Prob, a \mid Sol) = \sum_{a} p(Prob \mid a, \mathrm{Sol})\, p(a \mid Sol) \qquad (2)$$

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

$$p(Prob \mid Sol) = \frac{\varepsilon}{(l+1)^{m}} \prod_{j=1}^{m} \sum_{i=1}^{l} t(Prob_j \mid Sol_i) \qquad (3)$$

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 the EM algorithm described in the previous section.

Once the translation model is acquired, textual retrieval consists of evaluating equation (3) for each of the cases in the case base and selecting the one with the higher translation probability. For the sake of brevity, we do not address how language models can be incorporated in the estimation process. But we refer the reader to [6] for additional information.

5 Textual Adaptation using Translation Models

As opposed to structural CBR offering various approaches for adapting structured cases, adaptation of textual solutions has seldom been explored as a research topic. Most of the reuse approaches in structural CBR consist of modifying the values of well-structured features. In a textual setting, it reveals difficult to implement similar schemes as the passages to be modified in a solution cannot be determined a priori. Hence, prior to modifying a textual solution, one is required to determine the basic units of text to process, their relevance and their specificity.

We could envisage that some approaches, based on natural language processing techniques, be proposed for tackling some of the following tasks:

- Substitutional adaptation: to replace passages, either words or phrases, of a text to enhance its relevance. This can be decomposed in two sub-problems: determining the words or passages to be modified and selecting relevant replacement values. Examples are the personalisation of information entities in texts (ex. to modify the name of an individual) or the substitution of some ingredients in a recipe (as encountered in the *Computer Cooking Contest*[1]).
- Transformational adaptation: to change the structure of a text. We can assume that transformation is required because some of the sentences or passages of the text are not relevant to the new problem to be solved. Hence we have to determine which portions of the description should be removed to improve the relevance of the solution. This formulation corresponds to a text summarization task and can be accomplished using translation models.
- Compositional adaptation: to combine sentences and passages from various textual solutions. This task could be implemented using multi-document summarization techniques but this issue is not addressed in this paper.

Let's consider the second task related to transformational adaptation. According to our formulation, this would correspond to the identification of non-relevant portions and to reorganize the content of an antecedent solution by pruning the superfluous parts. We will assume to simplify the discussion that decisions are to be made on the

[1] http://www.wi2.uni-trier.de/eccbr08/index.php?task=ccc

relevance of the sentence. Preserving the integrity of sentences favours the coherence and the intelligibility of the text resulting from the pruning process. We also assume that a sentence pertains to a single theme. But these assumptions are not critical for the application of the technique we discuss in the next paragraphs.

In order to find the sentences of a solution that are relevant to the new problem to be solved, we need to execute the three following tasks:

- Segmentation: We break the solutions into individual sentences.
- Evaluation: We estimate the relevance of individual sentences with respect to the content of the problem;
- Selection: We choose the sentences that optimize the global relevance of the transformed solution.

A possible strategy for conducting this process is to select a subset of sentences that best covers the content of the new problem to solve. This processing of the relevance at the sentences level corresponds to a reduction of the text. In natural language processing, this type of summarization process is frequently referred to as *query-biased* [11] or *user-centered*. More specifically, it corresponds to the production of a condensed text based on the terms of the new problem. In this formulation, a problem indicates the focus of the user (what is being looked for) and the portions of text that are found in the summary should be in agreement with the statements of the problem.

As illustrated in Figure 3, the resulting solution *Sol'* can be produced by the deletion, from the original solution *Sol*, of sentences that should not be associated (or *aligned*) to the new problem *Q*.

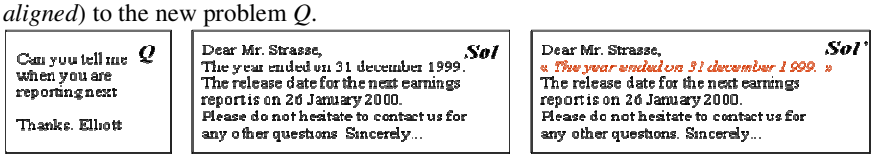

Fig. 3. Textual transformational adaptation as a summarization process (from [10]).

This adaptation process tries to search for a subset *Sol'* that covers most of the problem *Q*. In terms of probability, we are trying to find a condensed solution *Sol'* that maximize the following probability estimate:

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

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

$$Sol' = \arg\max_{Sol'} P(Q | Sol', Sol)\ P(Sol' | Sol) \tag{2}$$

$$\sim \arg\max_{Sol'} P(Q | Sol')\ P(Sol' | Sol)$$

Hence, this formulation of transformational adaptation suggests that the solution being recommended to the user of a CBR system is a compromise between the integrity of the past solution $P(Sol'|Sol)$ and a subset of the solution that best fits the new problem $P(Q|Sol')$. The latter expression can be modeled as a random withdrawal of terms from the original solution *Sol*. Some probability distributions (for instance

multinomial or hypergeometric) can be exploited for the evaluation of the resulting condensate text.

The expression *P(Q|Sol')* corresponds to the probability that a new problem *Q* is at the origin of a solution *Sol'*. This probability distribution can be modeled using a translation model indicating how problems can be transferred into solutions. We constructed this distribution in [10] by applying an IBM Model 1 to the case base of our application.

Once the global quality of a text can be assessed, a search algorithm is needed for the selection of the best subset of sentences. Exhaustive search is possible when solutions contain a few sentences (3-5). However, heuristic search are required when solutions are more voluminous. A best first search, using the translation probability of each individual sentence, provides good quality results in practice while reducing significantly the time needed to find a solution.

6 Discussion

We presented in this paper a short overview on how to apply translation models to the retrieval and adaptation phases of textual CBR systems. Various translation models are presented in the literature and we limited our coverage to word-based alignment models. We have tested the IBM Model 1 in some of our previous work on a case base of limited size with satisfactory results. However additional experiments would be required to evaluate the extent of these techniques on a large case base as the one proposed for the CCC contest. Also, due to their computational complexity, most of the alignment algorithms are not suitable for long descriptions. Hence sentence alignment would also be required to fragment the solution and problem descriptions. Additional work would also be required to assess the adequacy of phrase-based models in a CBR setting.

The problem of textual adaptation remains to be explored. Recently, Bentebibel *et al.* [12] proposed an approach for the compositional reuse of incident reports. As this approach relies strongly on manual efforts to structure the cases and to represent the knowledge contained in the reports, an automated approach for the compositional adaptation of textual cases is still an open issue.

Finally, as often experienced in textual CBR, evaluation of the results provided by statistical translation systems is a difficult task. The machine translation community has devoted considerable efforts to resolve this issue and to propose metrics (such as *BLEU* [13]) to assess the quality of text without human intervention. There might be some opportunities to transpose some of the lessons learned from this community to the evaluation of textual CBR systems.

References

1. Delany, S. J.; Bridge, D.: Textual Case-Based Reasoning for Spam Filtering - A Comparison of Feature-Based and Feature-Free Approaches, *Artificial Intelligence Review,* 26(1-2), pp. 75--87 (2006).

2. Burke, R.; Hammond, K.; Kulyukin, V.; Lytinen, S.; Tomuro, N.; Schoenberg, S.: Question Answering from Frequently-Asked Question Files: Experiences with the FAQ Finder System. *AI Magazine*, 18(2), pp. 57--66 (1997).
3. Lamontagne, L.; Lapalme, G.: "Applying Case-Based Reasoning to Email Response", *Proceedings of ICEIS-03*, Angers, France, pp. 115--123 (2003).
4. Massie, S.; Wiratunga, N.; Donati, A.; Vicari, E.: From Anomaly Reports to Cases. *Proceedings of ICCBR'07*, Springer (2007).
5. Brown, P. F.; Della Pietra, S. A.; Della Pietra, V. J.; and Mercer, R. L.: The mathematics of statistical machine translation: Parameter estimation, *Computational Linguistics*, vol. 19, no. 2, pp. 263--311 (1993).
6. Manning, C., Schütze, H.: *Foundations of Statistical Natural Language Processing*, MIT Press. Cambridge, MA (1999).
7. Jurafsky, D; Martin, J.: *Speech and Language Processing – An introduction to Natural Language Processing*, *Computational Linguistics and Speech Recognition*, 2nd edition, Prentice Hall (2008).
8. Koehn, P.; Och, F. J.; Marcu, D.: Statistical Phrase-Based Translation, *Proceedings of HLT/NAACL* (2003).
9. Lamontagne L., Langlais, P., Lapalme, G.: Using Statistical Models for the Retrieval of Fully-Textual Cases, *Proceedings of FLAIRS-2003*, pp.124--128, AAAI Press (2003).
10. Lamontagne, L.; Lapalme, G.; Textual Reuse for Email Response, *Advances in Case-Based Reasoning*, LNCS, vol. 3155, pp. 234--246, Springer-Verlag (2004).
11. Mittal, V.; Berger, A.: Query-relevant summarization using FAQs, *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics (ACL)* (2000).
12. Bentebibel R.; Despres S.: Using Compositional and Hierarchical Adaptation in the SAARA System. *3rd Textual Case Based Reasoning Workshop*, ECCBR'06 (2006).
13. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J.: "BLEU: a method for automatic evaluation of machine translation" in *ACL-2002: 40th Annual meeting of the Association for Computational Linguistics* pp. 311—318 (2002).