The Knowledge Engineering Review, Vol. 00:0, 1–00. © 2005, Cambridge University Press
DOI: 10.1017/S000000000000000 Printed in the United Kingdom

Textual case-based reasoning

ROSINA O. WEBER[1], KEVIN D. ASHLEY[2], STEFANIE BRÜNINGHAUS[2]

[1]*College of Information Science & Technology, Drexel University, Philadelphia PA 19104, USA*
E-mail: rosina.weber@drexel.edu
[2]*Learning Research and Development Center, School of Law, and* [3]*Intelligent Systems Program, University of Pittsburgh, Pittsburgh PA 15260, USA*
E-mail: ashley@pitt.edu
E-mail: bruninghaus@gmail.com

Abstract

This commentary provides a definition of textual case-based reasoning (TCBR) and surveys research contributions according to four research questions. We also describe how TCBR can be distinguished from text mining and information retrieval. We conclude with potential directions for TCBR research.

1 What is textual case-based reasoning?

Case-based reasoning (CBR) consists of comparing a new problem to previously solved cases in order to draw inferences about the problem and to guide decision making. Textual case-based reasoning (TCBR) is a subfield of CBR concerned with research and implementation on case-based reasoners where some or all of the knowledge sources are available in textual format. It aims to use these textual knowledge sources in an automated or semi-automated way for supporting problem solving through case comparison.

While the most well-known and widely used form of TCBR is the retrieval of textual cases, all phases of the CBR cycle are relevant for TCBR applications, and textual knowledge sources may impact both the reasoning cycle of a CBR system as well as the CBR system's design and development. The CBR cycle (Aamodt & Plaza, 1994) consists of four steps: when a new problem is submitted, the case-based reasoner *retrieves* similar cases. By adapting retrieved solutions, the *reuse* step determines a proposed solution, the *revision* step confirms the solution, and the *retain* step can incorporate the new case into the case base. As part of the design and development of CBR systems, some of the decisions to be made are how to identify problem solving experiences to populate the case base, what representation for cases to adopt, how to define the indexing vocabulary, which retrieval methods to adopt, and how to extract and represent reusable components. Textual knowledge sources may be involved in the development and reasoning cycle of TCBR systems in a variety of ways. For instance, the problems and cases themselves may be available as texts. The goals of the TCBR system might then be to retrieve the textual cases relevant to solving the textually-described problem, extract or highlight relevant passages in the textual cases, extract and assign indices to the textual cases so that they can be retrieved in the future, or to use the textual cases to reason interpretively about a problem. Alternatively, a TCBR system may perform only some of these tasks or employ textual knowledge sources in some other way to support problem solving with cases.

In the next section, we will identify four major research questions addressed in TCBR and survey contributions relating to them. In Section 2.2, we highlight (Brüninghaus & Ashley, 1999), which is a prototypical research contribution to TCBR. Finally, we clarify some distinctions between TCBR and other text-oriented methods and identify research opportunities in TCBR.

2 An overview of TCBR research

Over the years, there has been significant progress addressing the threshold challenge facing TCBR, how to bring textual knowledge sources to bear in supporting reasoning with cases. Specifically, the research has addressed four questions: (1) how to assess similarity between

textually represented cases; (2) how to map from texts to structured case representations; (3) how to adapt textual cases; and (4) how to automatically generate representations for TCBR. These focal questions of TCBR will organize this survey of the state of the art.

2.1 Similarity between textually represented cases

Some of the pioneer work in TCBR demonstrated how CBR techniques can be applied to retrieval tasks. These approaches do not rely on a symbolic representation of cases but compare these cases as text tokens using a variety of techniques adapted from information retrieval (IR). They achieve a richer notion of case similarity by supplementing the textual comparisons with basic linguistic techniques and methods that take the meaning of words into account.

Burke *et al.* (1997) developed FAQ-Finder, a question-answering system. Given as input a typed question, it retrieves textual answers from Usenet FAQ files, which contain frequently asked questions with answers. Conceptually, each of the question-answer pairs is treated as problem and solution in a CBR framework. FAQ-Finder uses techniques that combine statistical and semantic knowledge. It starts with a standard IR approach based on the vector space model, where cases are compared as term vectors with weights based on a term's frequency in the case versus in the corpus. In addition, FAQ-Finder includes a semantic definition of similarity between words, which is based on the concept hierarchy in WordNet (Fellbaum, 1998). An evaluation showed that adding semantic information led to performance improvements. FAQ-Finder was one of the first TCBR implementations that demonstrated the benefits from incorporating background knowledge.

Lenz and Burkhard (1997) took a different approach in FAllQ, another question-answering system that compares textual cases through the meanings of terms. Cases consist of a question text, a list of attributes, and the answer text. The program processed the free text components to identify Information Entities (IE), which are indexing concepts that may occur in text in different forms. This approach requires some domain-specific knowledge engineering to identify task-specific terms, which may include product names or physical units. FAllQ's similarity assessment checks word similarity using two lexical sources: a manually constructed domain-specific ontology and a generic thesaurus. Case Retrieval Nets, which support FAllQ's retrieval strategy, represent the case base as a network of IE nodes where similarity arcs connect nodes with similar meaning. Retrieval is performed by propagating activation through this network.

Wilson and Bradshaw (2000) investigated cases that required mixed representations including both textual and non-textual features. They used the IR term vector space model to assess individual similarities between the textual features and integrated them into case similarity assessment techniques for the non-textual features.

2.2 From texts to structured case representations

Another group of projects focused on developing methods to map textually expressed cases into the kinds of structured representations used in CBR systems. SPIRE (Rissland & Daniels, 1996) is a hybrid CBR/IR system that supports humans who perform case-based legal research. It retrieves relevant case texts from the full-text database and highlights the most relevant portions given the target case. A human might use the system to index legal cases from a large full-text database for use in a CBR system. She would describe a problem in terms of the structured CBR representation. The CBR system retrieves the most relevant cases for the user's problem from its manually indexed case base, which contains only a small portion of the universe of cases that are stored in the large full-text database. The original text of the most relevant cases is given to the relevance feedback module of an IR system, which retrieves the most similar case texts. These cases are likely to be relevant to the user's problem, but still need to be indexed in terms of the CBR representation. In a second step, SPIRE addresses this problem. For each of the features in the CBR case representation, SPIRE has a small collection of textual excerpts associated with the feature. A subset of these textual excerpts is used as a query to the IR system for retrieving similar passages in the case texts retrieved in the first step. For each CBR index,

the IR system presents a ranked list of candidate passages to the human user, which will significantly decrease the amount of work required to add these new cases to the case base.

Weber *et al.* (1998) introduced a semi-automated approach to populate case templates from textual documents. The method is domain specific and requires knowledge engineering to elicit from domain experts which attributes are relevant in the target domain, how to find them, and the variations in which they might occur in the domain-specific texts. This knowledge is then used to feed template mining methods that extract the feature values from text to populate the cases. Template mining bypasses natural language processing (NLP) by relying on the structure identified through linguistic patterns found in text. One feature, which was sensitive to negation, required NLP; all other features were successfully extracted without NLP. The case templates include both indexing and reusable features. These features are shown to the user in the implemented system, named PRUDENTIA, to help users manually select texts and reuse them.

Brüninghaus and Ashley (1999) applied text classifiers to automate the mapping from texts to structured case representations. The indexing concepts were Factors, which are stereotypical fact patterns that tend to strengthen or weaken a legal claim. The case texts, factual descriptions of legal disputes, were represented as a bag-of-words (BOW), the representation commonly used in IR and text learning. In SMILE, sentences, rather than entire case texts, were used as positive and negative instances from which text classifiers for each Factor were learned. In the implementation, background knowledge in the form of a domain specific thesaurus was used to expand words into synonym sets. An empirical evaluation showed that this technique led to performance improvements. This paper received the best paper award at the 1999 International Conference on CBR, and it was recognized as the most significant contribution in TCBR at the 2003 New Zealand Workshop on CBR.

Building upon the initial results with SMILE, Brüninghaus and Ashley (2001) introduced two innovations to improve on the BOW representation. To generalize from the training examples, they proposed replacing case-specific names and instances in the sentences with their roles in the case (e.g., the roles of "plaintiff" and "defendant" in a lawsuit). They also introduced Propositional Patterns (ProPs), syntax-based multi-word features that capture information about who did what to whom. ProPs are derived with information extraction tools to include both the words and syntactic patterns in the examples, including subject-verb, verb-object, verb-prepositional phrase and verb-adjective. In addition, ProPs include certain semantic information about negation and selected adjective labels. While ProPs are not a deep representation of the text, they are more expressive than BOW, offering potential improvements in performance for assigning the Factors. Because some of the most relevant Factors correspond to actions, it is important to capture a sense of who did what to whom.

Brüninghaus and Ashley (2005) completed the loop by demonstrating how the SMILE+IBP framework can use the representation automatically generated by SMILE as input for the interpretive CBR program IBP to reason about textually described problems. Specifically, SMILE+IBP makes a case-based prediction for the outcome of a legal case from a brief textual summary of the facts. An evaluation demonstrated that SMILE+IBP achieved better prediction performance than some reasonable alternatives. The experiments also showed that the role replacements and ProPs in SMILE improved significantly on the BOW representation. This work on integrating automated indexing and reasoning about case texts demonstrates the value of applying NLP to improve a TCBR system's performance. Likewise, Mott *et al.* (2005) showed that syntax analysis can improve performance for textual case retrieval tasks.

2.3 Adapting textual cases

The concept of adapting cases for reuse, a distinguishing feature of CBR, is relevant for TCBR as well. Lamontagne and Lapalme (2004) introduced a novel approach for adapting a solution from a retrieved textual case to solve the target problem. Cases are emails and thus the goal of the system is to generate a response for an incoming message, a request. The case base contains past messages organized in cases that consist of request and response. After retrieving the best case for a new target message, their approach examines the response of the retrieved case and

categorizes its sentences to identify which portions are reusable. It then modifies and adapts the retrieved response to address the unmatched aspects of the request. This research pushes the limits of TCBR because it exemplifies reasoning with textual cases for problem solving with adaptation.

2.4 Towards automatically generating representations for TCBR

A relatively recent line of research in TCBR focuses on automated case representation and retrieval methods. These projects introduce novel techniques for enabling a program to induce or otherwise discover general knowledge for representing textual cases. While these approaches may not perform problem solving via case comparison, they are promising for TCBR research.

Wiratunga *et al.* (2004) introduced a fully automated method for extracting predictive features to represent textual cases. The approach included feature selection with boosting and association rule induction to discover semantic relations between words. Subsequently, Wiratunga *et al*. (2005) extended their approach to generate propositional clauses that represent logical combinations of keywords. The resulting representation of textual cases consists of interpretable features such as the clause "intelligent" $\vee$ "algorithm" $\vee$ ("grant" $\wedge$ "application").

Cunningham *et al*. (2004) investigated the automated construction of graphs to represent textual cases for TCBR and overcome the limitations of a BOW. The approach represents textual cases as graphs; the nodes are words, and arcs are added between adjacent words. It preserves word order and can capture important features like negation. The similarity between cases is calculated with graph-distance methods. While the algorithms are domain-independent, domain-specific knowledge is used implicitly by prioritizing the most relevant terms in the process. The approach's inability to distinguish problems from solutions has posed a limitation for case reuse. This problem is being addressed in ongoing research (Proctor *et al*., 2005).

Patterson *et al.* (2005) presented SOPHIA, a text clustering approach that does not require labeled data. Using term distributions, SOPHIA builds themes, which are groups of words that appear in similar documents. Clusters are semantically similar texts that share themes. These clusters succeed in distinguishing meanings of expressions even if they are polysemous. Also, SOPHIA can cluster texts with similar meanings even if they have different terminology. Although the authors have not yet harnessed SOPHIA to solve problems by case comparison, the ability to identify and represent clusters of semantically related texts is promising for TCBR.

Gupta and Aha (2004) proposed a deep natural language understanding approach for TCBR that derives a first-order representation of the case texts. The envisioned system will also identify relevant attributes for the case representation dynamically. While intriguing, this approach poses extreme knowledge representation and engineering challenges. Because the proposed methods go beyond the currently available technology, considerable research will be required before an implementation becomes possible.

3 Distinguishing TCBR's contributions and challenges

The applications of TCBR are related to a number of other research areas concerned with textual documents like IR and text mining. Text mining focuses on the discovery of information and knowledge from unstructured documents, whereas IR identifies documents that satisfy a user's information needs expressed through a query (Baeza-Yates & Ribeiro-Neto, 1999). The techniques developed in IR and text mining are largely domain- and task-independent, with a focus on general-purpose systems.

The most important distinction between these areas and TCBR is the system's goal. While TCBR may apply text-oriented techniques like document clustering or information extraction, similar to and sometimes even adapted from IR or text mining, its focus is on reasoning and problem solving with cases. This goes well beyond the fairly basic and well-defined tasks in IR. IR and text mining methods are by their nature general-purpose; they do not incorporate background domain knowledge or consider how the information will be used. Like

IR systems, successful question-answering systems are not intended for a specific problem solving task, even though they may employ syntactic and semantic information.

In contrast, TCBR attempts to leverage both task- and domain-specific knowledge. IR researchers might dismiss such techniques as *ad hoc*, but it allows textual CBR systems to "eschew flexibility and generality for precision and utility for a given group of users" (Burke, 1998). Text-oriented comparison methods, originally designed for IR tasks, do not embed such knowledge and in particular cannot guarantee finding similar cases with adaptable solutions. On the other hand, TCBR requires a comparatively well defined problem solving task and relationship between queries and cases.

This discussion also helps to identify future directions for TCBR. TCBR is a unique research area whose challenges come from the combination of textual documents and the problem solving and reasoning that sets CBR apart from other AI methodologies. In its brief history, TCBR projects have developed new methods for retrieving and representing cases, adapting and improving the text processing techniques used in IR and text mining. Recent TCBR systems have pushed the capabilities of TCBR further by implementing more steps of the CBR cycle (Lamontagne & Lapalme, 2004; Brüninghaus & Ashley, 2005). Open challenges include increasing accuracy, more complete automation, and the ability to process more complex texts in a wide variety of domains.

References

Aamodt, A. & Plaza, E. 1994, "Case-based reasoning: foundational issues, methodological variations, and system approaches" *AI Communications*, 7 (1), pp. 39-59.

Baeza-Yates R. & Ribeiro-Neto, B. 1999, *Modern Information Retrieval*. Addison-Wesley, Cambridge, MA.

Brüninghaus, S. & Ashley, K. D. 2005, "Reasoning with Textual Cases" in Muñoz-Avila, H. & Ricci, F. eds. *Case-Based Reasoning Research and Development* (LNAI 3620), Springer, Berlin, pp. 137-151.

Brüninghaus, S. & Ashley, K. D. 2001, "The Role of Information Extraction for Textual CBR" in Aha, D. W. & Watson, I. eds. *Case-Based Reasoning Research and Development* (LNAI 2080), Springer, Berlin, pp. 74-89.

Brüninghaus, S. & Ashley, K. D. 1999, "Bootstrapping Case Base Development with Annotated Case Summaries" in Althoff, K. D., Bergmann, R. & Branting, L. K. eds. *Case-Based Reasoning Research and Development* (LNAI 1650), Springer, Berlin, pp. 59-73.

Burke, R. D., Hammond, K. J., Kulyukin, V., Lytinen, S. L., Tomuro, N. & Schoenberg, S. 1997, "Question answering from frequently-asked questions files: Experiences with the FAQ Finder system" *AI Magazine*, 18(1), pp. 57-66.

Burke, R. 1998, "Surveying Opportunities for Textual CBR" in *Textual Case-Based Reasoning: Papers from the AAAI-98 Workshop*, In Lenz, M. & Ashley,K. D. eds., AAAI Press, Menlo Park, CA, pp. 13-18.

Cunningham, C., Weber, R., Proctor, J. M., Fowler, C. & Murphy, M. 2004, "Investigating Graphs in Textual Case-Based Reasoning" in Funk, P. & González Calero, P. A. eds. *Advances in Case-Based Reasoning* (LNAI 3155), Springer, Berlin, pp. 573–586.

Fellbaum, C. ed. 1998, *WordNet: An Electronic Lexical Database*, The MIT-Press, Cambridge, MA.

Gupta, K. M. & Aha, D. W. 2004, "Towards acquiring case indexing taxonomies from text" in Barr, V. & Zdravko, M. eds. *Proceedings of the Seventeenth Annual Conference of the International Florida Artificial Intelligence Research Society*, AAAI Press, Menlo Park, CA, pp. 172-177.

Lamontagne, L. & Lapalme, G. 2004, "Textual Reuse for Email Response" in Funk, P. & González Calero, P. A. eds. *Advances in Case-Based Reasoning* (LNAI 3155), Springer, Berlin, pp. 242–256.

Lenz, M. & Burkhard H. D. 1997, "CBR for Document Retrieval - The FAllQ Project" Leake, D. B. & Plaza, E. eds. *Case-Based Reasoning Research and Development* (LNAI 1266), Springer, Berlin, pp. 84-93.

Mott, B., Lester, J. & Branting L. K. 2005, "The Role of Syntactic Analysis in Textual Case Retrieval" in Weber, R. & Branting, L.K. eds.*Proceedings of the Textual Case-Based Reasoning Workshop*, Chicago, IL, pp.120-127.

Patterson, D., Dobrynin, V., Galushka, M. & Rooney, N. 2005, "SOPHIA: A Novel Approach for Textual Case Based Reasoning" in Kaelbling, L.P. ed. *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*, Morgan Kaufman Publisher, San Mateo, CA.

Proctor, J., Waldstein, I. & Weber, R. 2005, “Identifying Facts for TCBR” in Weber, R. & Branting, L. K. eds. *Proceedings of the Textual Case-Based Reasoning Workshop*, Chicago, IL, pp. 150-159.

Rissland, E. L. & Daniels, J. J. 1996, "The Synergistic Application of CBR to IR" *Artificial Intelligence Review*, 10 (5-6), pp. 441-475.

Weber, R., Martins, A. & Barcia, R. 1998, “On legal texts and cases” in Lenz, M. & Ashley, K. D. eds. *Textual Case-Based Reasoning: Papers from the AAAI-98 Workshop*, AAAI Press, Menlo Park, CA, pp. 40-50.

Wilson, D. C. & Bradshaw, S. 2000, “CBR Textuality”, *Expert Update*, 3(1). pp. 28-37.

Wiratunga, N., Koychev, I. & Massie, S., 2004, “Feature Selection and Generalisation for Retrieval of Textual Cases” in Funk, P. & González Calero, P. A. eds. *Advances in Case-Based Reasoning* (LNAI 3155), Springer, Berlin, pp. 806-820.

Wiratunga, N., Lothian, R., Chakraborti, S. & Koychev, I. 2005, “Textual Feature Construction from Keywords” in Proceedings of the Textual Case-Based Reasoning Workshop, Chicago, IL, pp. 110-119.