Towards a case-based reasoning approach to analyze road accidents

Valentina Ceausu[1]
Sylvie Després[1]

René Descartes University
45 rue des Saints Pres
75270 Paris cedex 06
valentina.ceausu@math-info.univ-paris5.fr
sd@math-info.univ-paris5.Fr

Abstract. In this paper the prototype of a system designed to analyze road accidents is presented. The analysis is carried out in order to recognize within accident reports general mechanisms of road accidents that represent prototypes of road accidents. Case Based Reasoning (CBR) is the chosen problem solving paradigm. Natural language documents and semi-structured documents are used to build the cases of our system, which creates a difficulty. To cope with this difficulty we propose approaches integrating semantic resources. Hence, an ontology of accidentology and a terminology of road accidents are used to build cases. The alignment of two resources supports the retrieval process. A data processing model, based on models of accidentology, is proposed to represent the cases of the system. This paper presents the architecture of ACCTOS (ACCident TO Scenarios), a case based reasoning system prototype. The model to represent the cases is introduced and the phases of the case based reasoning cycle are detailed.

1 Introduction

In this paper the prototype of a system designed to analyze road accidents is presented. The analysis is carried out in order to recognize, within accident reports, general mechanisms of road accidents that represent prototypes of road accidents.
Case based reasoning is the chosen problem solving paradigm. Cased based reasoning solves a new problem by re-using a collection of already solved problem. The problem to be solved is called the target case. The collection of already solved problems make up the case base, an important feature of any case based reasoning system. The reasoning cycle of a case based reasoning system is composed of phases aiming to: (i)create the target case; (ii) retrieve cases of the case base which are similar to the target case; (iii) adapt solutions of some of these cases in in order to propose a solution for the target case.
Natural language documents and semi-structured documents are used to build our system cases. To cope with the difficulty of natural language, we proposed

approaches integrating semantic resources. An ontology of accidentology and a terminology of road accidents are used to build descriptions of cases. The alignment of two resources supports the retrieval process. Based on accidentology models, a data-processing model is proposed to represent the cases of the system.
The outline of this paper is as follows: first, the architecture of ACCTOS (ACCident TO Scenarios) is presented and the model proposed to represent cases of the system is introduced. Then, phases of the case based reasoning cycle are detailed. Finally, conclusions are drawn and future works are suggested.

2 Architecture and resources of the system

To present the architecture, we use a division into modules, where each of the module addresses a different phase of the reasoning cycle (see Fig. 1).

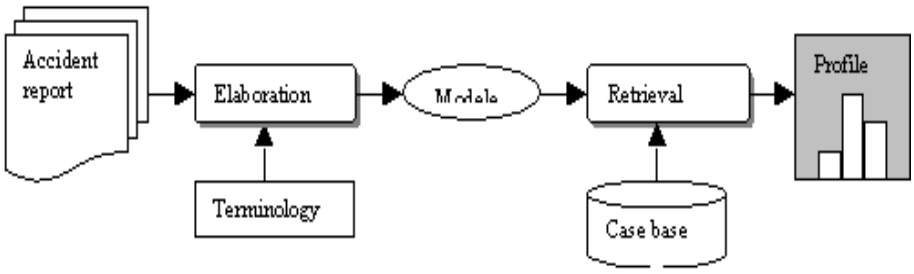

Fig. 1. System architecture

Resources of the system
ACCTOS exploits two types of documents -accident reports and accident scenarios -to create cases.
Accident reports are documents written by the police. They include structured paragraphs describing An accident actors and context and natural language paragraphs explaining what happened in the accident (written with the help of witnesses, people involved in the accident or policemen). Accident scenarios are documents created by road safety researchers. They are prototypes of road accidents and present in a general way facts and causal relations between different phases leading to collision. An accident scenario describes an accident as a sequence of four situations (or phases): the driving situation, the accident situation, the emergency situation and the shock situation. Prevention measures aiming to improve road safety are provided for each accident scenario. A first study led by the department *Mechanisms of accidents* of INRETS (Institut national de recherche sur les Transport et leur Sécurité) established a first collection of accident scenarios involving pedestrians. These scenarios and

assigned proposals will be used to build the case base.
The input of the system is a set of reports of accident that occurred on the same road section. Accidents are analyzed from electronic accident reports. The *PACTOL* tool (Centre d'Etudes Techniques de l'Equipement de Rouen) made the reports anonymous. An electronic accident report is a semi-structured document containing structured paragraphs and natural language paragraphs. Structured paragraphs contain variables describing the accident. The variables correspond to humans and vehicles involved in the accident.The accident context is also specified by variables. Text paragraphs describe what happened in the accident according to several points of view: the police (synthesis of the facts), the people involved (declarations) and the witnesses.
From each accident report, a model is built by the *Elaboration* module. This model is used by the *Retrieval* module of the system in order to query the case base. The initial case based of ACCTOS is created from accident scenarios. As a result, correspondences between the initial accident report and accident scenarios are established. A correspondence is constituted by an assignment (*accident report, accident scenario*) and a trust assessment.
The output of the system is a profile of scenarios. A profile of scenarios is composed of several scenarios, where a coefficient is assigned to each scenario of the profile, reflecting its weighting within the profile.
The first module implements the authoring the case phase. Retrieval phase is done by the second module. Further on the model proposed to represent cases of the system and phases of case based reasoning cycle are presented.

3 Description of cases

A data-processing model is proposed, based on accidentology models (see [5]), to represent cases of the system. A case is described by two types of elements: global variables and agents.
Global variables specify the number of agents involved in an accident, the environment in which the accident occurred - such as main road or secondary road - and the accident context (by day, in intersection, etc.).
A human involved in an accident and his vehicle represent an agent (see tab.1). This representation allows us to cope with difficulties related to metonymy between the human involved in the accident and his vehicle. It also allows us to treat the particular case of pedestrians. Each agent is defined by two components - human and vehicle - and by his evolution in the accident. A domain term (*ie: driver, car*) and attributes (*ie: age*) are assigned to each component of an agent. Agent evolution is specified by a set of relations describing interactions: between an agents' components; between an agent and other agents involved in the accident.

Table 1. Components of an agent

Agent	Human	Vehicle	Attributes	Evolution
Agent 1	Pedestrian	no vehicle	age: 60	crossing; running
Agent 2	Driver	Car	age: 35	moving; turning to

4 Authoring the case

The scope of this phase is to create the problem to be solved, also called the target case.The model presented above is used to represente the target case. Each target case is created from an accident report. Both the structured and the natural language paragraphs of an accident report are exploited to create the target case.

Environnement identification

An accident report is a semi-structured document. Data about people and vehicles involved in an accident and about the environment and context of the accident are stored in specific structures. Based on these structures we have designed automatic procedures to retrieve valuable information.

Identification of agents

To describe agents involved in accidents we need to :

- Identify the terms assigned to their components;
- Identify the values of their attributes;
- Identify the agents'evolution;

Terms of components and values of attributes are identified automatically thanks to the accident report structure.
Agents' evolution is identified thanks to natural language paragraphs of accident reports: declarations, testimonies, police syntheses. Agent evolution is expressed by a set of domain verbs identified within these paragraphs.
Text mining techniques and a terminology of road accidents are used jointly to identify the evolution of each agent.
A terminology represents terms of a given field and relations between those terms. Relations are expressed by verbs and, usually, accepts two arguments: *Relation(domain, range)*, where *Relation* is a verb of the field, and *domain* and *range* are terms of the field.
For instance, *diriger-vers(véhicule, direction)* is a relation of the domain. We used a terminology created from 250 reports of accidents that occurred in and around the Lille region. This terminology is expressed in OWL (see [15]).
Text mining techniques are also employed to identify agent evolution. An approach based on information extraction using pre-defined patterns is adopted.

We used lexical patterns to extract information. A lexical pattern is a set of lexical categories. For example *Noun, Noun* or *Verb, Preposition, Noun* are lexical patterns. In order to identify instances of patterns, natural language paragraphs are tagged using TreeTagger ([10]). A pattern recognition algorithm (see [2]) allows us to identify associations of words matching predefined patterns. The output of this algorithm is shown below :

Lexical Patterns and Corresponding word regroupings :
Noun, Preposition, Noun: groupe de piéton (group of pedestrians)
Noun, Preposition, Adjective: trottoir de droite (right hand side pavement)
Verb, Preposition, Noun: diriger vers place (direct to square)

We defined a set of verbal patterns able to highlight relations of the domain. A set R of verbal relations is extracted. Instances of those patterns could represent relations of the field, such as *diriger-vers (direct to)*, but also meaningless word regroupings, such as *diriger 306 (direct 306)*. They need to be validated and attached to agents of the accident. To do so, each agent $a(t_h, t_v)$, having t_h and t_v as components, queries the terminology in order to identify relations that have one of his components as arguments. The result is:

$$R_{agent}(t_h, t_v) = R_{resource}(t_h) \bigcup R_{resource}(t_v)$$

where $R_{resource}(t_h)$ and $R_{resource}(t_v)$ are relations of terminology having t_h, respectively t_v as arguments. By intersecting R_{agent} and R the evolution of an agent is identified as:

$$Evolution_{agent} = R_{agent} \bigcap R$$

Relations of R that are not modeled by the terminology are ignored. For each agent, the evolution is identified as a set of verbal relations extracted from the accident report and validated by the terminology of road accidents.

5 Building the initial case base

The case base is an important feature of a case based reasoning system. the case base is composed of couples $Problem, Solution$, that are called source cases.
A set of accident scenarios is used to build the initial case base of the system. The accident scenario represents the *Problem*; measures of preventions assigned to the scenario represent the *Solution*.
An **ontology of accidentogoly**(see [4]) supports descriptions of source cases. This ontology was built from expert knowledge, texts of the field and accident scenarios. It models the concepts of the field and the relations that hold between them. Ontology concepts are structured in thre main classes: the human, the vehicle and the environment. A domain term and attributes are assigned

to each concept. Concepts are connected by different types of relations. IS-A relations build the hierarchy of domain concepts. Verbal relations that describe interactions between concepts are also modeled.
An editor of scenarios was developed to build source cases. The editor integrates the ontology of accidentology. It allows users to describe each accident scenario by choosing the appropriate concepts and relations of the ontology. The editor also allows users to assign to each concept or relation a coefficient indicating its importance. Importance coefficients are established thanks to linguistic markers. Homogeneous descriptions of cases are created by integrating the ontology.

6 Retrieval process

The retrieval process aims to retrieve source cases similar to the target case. Already solved problems similar to the target case are identified. Therefore, a solution to the target case can be proposed by adapting solutions of those problems. We propose a retrieval approach supported by the alignment of two semantic resources : the terminology and the ontology.
Ontology alignment can be described as follows: given two resources each describing a set of discrete entities (which can be concepts, relations, etc.), find correspondences that hold between these entities. In our case, a function $Sim(E_o, E_t)$ is used allowing us to estimate similarity between entities of the ontology, E_o, and entities of the terminology, E_t, where an entity could be either a concept or a relation. Based on this, for T, a target case, two steps are needed to retrieve similar source cases.
(1) The first step is based on case base indexation. Global variables are used to index the case base. The values of the global variables of the target case are taken into account to identify a set of source cases. The result is a set of source cases having the same context as the target case and involving the same number of agents.
(2)A voting process is used to improve this first selection. The vote is done by each target case agent to express the degree of resemblance between himself and agents of a source case. A note is given by each target case agent to every source case. This note is given by taking into accounts agents' components and theirs evolution. A first similarity measure proposed is given by:

$Sim(a_i, a_j) = SimComponent(a_i, a_j) + SimEvolution(a_i, a_j)$
if $SimComponent(a_i, a_j) \neq 0$, otherwise $Sim(a_i, a_j) = 0$

where a_i is an agent of the target case and a_j is an agent of a source case.

$SimComponent(a_i, a_j)$ expresses the similarity among the agents taking into account component similarities :

$$SimComponent(a_i, a_j) = ch_j * sim(H_i, H_j) + cv_j * sim(V_i, V_j)$$

, where ch_j and cv_j are importance coefficients established for the source case, and values of $sim(H_i, H_j)$ and $sim(V_i, V_j)$ are given by the alignment of the two resources.
Evolution similarity expresses resemblances between of two agents' evolutions :

$$SimEvolution(a_i, a_j) = \frac{\sum_r c_r * sim(rSource_r, rTarget_r)}{\sum_r c_r}$$

Coefficients c_r expresses the importance of $rSource_r$ relation for the considered source case. Values of $sim(rSource_r, rTarget_r)$ are given by alignment of the two resources.
Each agent of the target case evaluates his resemblance to agents of the source case by using the presented approach. A similarity vector is obtained. The note $note_i$ given by the $agent_i$ to the source case is the maximum value of this similarity vector. Based on notes given by agents, the similarity between he target case and a source case is estimated by the average value:

$$Sim(target, source) = \frac{\sum_{i=1}^{N_a} note_i}{N_a}$$

where $note_i$ is the note granted by the agent $agent_i$, and N_a is the number of agents of the considered target case. Indexing the case base allows a fast identification of source cases that are similar to the target case. By voting, the most similar cases are selected among the cases retrieved by the first selection. The retrieval process is driven by the description of source cases whose importance coefficients are taken into account by similarity measures.

7 Conclusion and future work

This paper presents the prototype of a system designed to analyze road accidents. Case based reasoning is the adopted problem solving paradigm. Cases of the system are created from semi-structured documents provided by two different communities : accident reports written by the police and accident scenarios created by road safety researchers. Semantic resources are used to cope with heterogeneity and natural language representations. A terminology of road accidents supports the authoring the case phase. Description of source cases is supported by the ontology of road accidents. The alignment of a road accident terminology and ontology enables the retrieval process.
This system is under development. There now remains to implement the proposed approaches, evaluate the system and make it better.
To do so, a few lines of research are already considered, as for example : enriching the text mining techniques so as to improve the authoring the case phase ant obtain more precise descriptions of target cases.

References

1. R. Bergmann , On the use of Taxonomies for Representing Case Features and Local Similarity Measures, 6th German Workshop on Case-Based Reasoning, 1998
2. V. Ceausu and S. Desprès , Towards a Text Mining Driven Approach for Terminology Construction, 7th International conference on Terminology and Knowledge Engineering, 2005
3. Cherfi and A. Napoli and Y. Toussaint, Vers une méthodologie de fouille de textes s'appuyant sur l'extraction de motifs fréquents et de règles d'association, onférence francophone sur l'apprentissage automatique, 2003
4. S. Desprès, Contribution la conception de méthodes et d'outils pour la gestion des connaissances, Université René Descartes, Paris, October, 2002
5. D. Fleury, Sécurité et urbanisme. La prise en compte de la sécurité routière dans l'aménagement urbain, Presses de l'Ecole Nationale des Ponts et chaussées, Paris, November, 1998
6. K.M. Gupta and D.W. Aha and N. Sandhu, Exploiting taxonomic and causal relations in conversational case retrieval, Sixth European Conference on Case-Based Reasoning, 2002
7. M. Klein, Combining and relating ontologies: an analysis of problems solutions, Workshop on ontologies and Information sharing, IJCAI , 2001
8. U. Hahn and K. Schnattinger, Towards text knowledge engineering, American Association for Artificial Intelligence Conference , 1998
9. P. Seguela, Adaptation semi-automatique d'une base de marqueurs de relations sémantiques sur des corpus spécialisés, Terminologie et Intelligence Artificielle, 1999
10. , H. Schmid, Probabilistic part-of-speech tagging using decision trees, International Conference on New Methods in Language Processing, 1994
11. B. Smyth and P. McClave, Similarity vs. diversity, 4th International Conference on Case-Based Reasoning, 1994
12. R. Weber and D.W. Aha and N. Sandhu and H. Munoz-Avila, A textual case-based reasoning framework for knowledge management applications, 4th Ninth German Workshop on Case-Based Reasoning, 2001
13. N. Wiratunga and S. Craw and S. Massie, Index Driven Selective Sampling for CBR, 5th International Conference on Case-Based Reasoning, 2003
14. N. Wiratunga and I. Koychev and S. Massie, Feature Selection and Generalisation for Retrieval of Textual Cases, 7th European Conference on Case-Based Reasoning, 2004
15. World Wide Web Consortium (W3C), OWL - Web Ontology Language,