Risk Analysis and Prevention in Procedures: extraction and preliminary results

[Extended Abstract]

Patrick Saint-Dizier
IRIT-CNRS 118 route de Narbonne 31062 Toulouse cedex France
stdizier@irit.fr

1. PROBLEMATICS

Maintenance operations as well as production launches are essentially based on procedures which describe how to install and use a product and how to maintain it. Due to the complexity of to-day's equipements, and to the complexity of their interactions it is difficult to maintain up-to-date documentations. These procedural documents become more and more complex, even if simplified language constraints and revision scenarios are imposed. According to several analysis, out of 377 technicians working in different domains, 45% of them indicate that they have identified major errors in maintenance documents. About 75% indicate that there are major gaps (missing instructions) or obscure or incomplete instructions, and 78% admit that often need help because they feel they are not operating the right way. We are all confronted to situations where we wish to follow instructions (DIY, software installation, etc.) with pictures, diagrams, etc. and that these are not understandable, have obvious gaps or do not correspond to the situation at stake. In some industrial areas, such difficulties are common and lead to accidents (aeronautics, nuclear energy, health, etc.). Risk analysis and prevention are therefore a major concern.

2. A DOMAIN-INDEPENDENT ANALYSIS OF RISKY SITUATIONS FROM TEXT ANALYSIS

Procedural texts consist of a sequence of instructions, designed with some accuracy in order to reach a goal (e.g. assemble a computer) [2, 3, 7, 8]. Procedural texts are complex structures, they often exhibit a quite complex rational (the goal-instructions) and 'irrational' structure which is mainly composed of advice, conditions, preferences, evaluations, user stimulations, etc. They form what we call the explanation structure [6], which motivates and justifies [4] the goal-instructions structure, viewed as the backbone of procedural texts. A number of these elements are forms of argumentation [1, 9], they appear to be very useful, sometimes as important as instructions: they provide a strong and essential internal cohesion and coherence to procedural texts. They also indicate, among other things, the difficulties, the risks to avoid, and the consequences on the target goal of an incorrect or incomplete execution of the associated instruction.

This is realized in the <TextCoop> [3] project, where a number of structures are tagged. An example, in readable form, from didactics, is given hereafter.

2.1 Measuring the intrinsic difficulty rate d of an instruction

It is of much interest to be able to measure the inherent complexity or difficulty of an instruction. This notion obviously depends on the reader profile. Nevertheless, we think that some linguistic features introduce some inherent difficulties in any situation.

The most frequently encountered parameters are, informally:
- presence of 'complex' manners (e. g. *very slowly*), by complex we mean either a manner which is inherently difficult to realize or a manner reinforced by an adverb of intensity,
- technical complexity of the verb or the verb compound used: if most instructions include a verb which is quite simple, some exhibit quite technical verbs, metaphorical uses, or verbs applied to unexpected situations, for which an elaboration is needed.
- duration of execution as specified in the instruction (the longer the more difficult),
- synchronization between actions, in particular in instructional compounds,
- uncommon tools, or uncommon uses of basic tools (*open the box with a sharp knife*) however this is quite difficult to chracaterize, besides statistical analysis (e.g. via bootstrapping on the net),
- presence of evaluation statements or resulting states, for example to indicate the termination of the action (*as soon as the sauce turns brown add flour*).

For some of these criteria, some application-dependent knowledge linguistic resources are needed: some lexical data, basic ontological data, and a few business rules. These observations allow us to introduce a very preliminary measure of complexity. To be able to have an indicative evaluation, each of the points above counts for 1, independently of its importance or strength in the text. Complexity c therefore

[$_{procedure}$ [$_{purpose}$ Writing a paper: [$_{elaboration}$ Read light sources, then thorough]]
[$_{assumption/circumstance}$ Assuming you've been given a topic,]
[$_{circumstance}$ When you conduct research], move from light to thorough resources [$_{purpose}$ to make sure you're moving in the right direction].
Begin by doing searches on the Internet about your topic [$_{purpose}$ to familiarize yourself with the basic issues;]
[$_{temporal-sequence}$ then] move to more thorough research on the Academic Databases;
[$_{temporal-sequence}$ finally], probe the depths of the issue by burying yourself in the library.
[$_{warning}$ Make sure that despite beginning on the Internet, you don't simply end there.
[$_{elaboration}$ A research paper using only Internet sources is a weak paper, [$_{consequence}$ which puts you at a disadvantage...]]]
While the Internet should never be your only source of information, [$_{contrast}$ it would be ridiculous not to utilize its vast sources of information. [$_{advice}$ You should use the Internet to acquaint yourself with the topic more before you dig into more academic texts.]]]

Figure 1: The explanation structure annotated in a procedure

ranges from 0 to 6. The complexity rate d_i of instruction i is $c/6$ to keep it in [0,1].

2.2 Measuring the expliciteness rate t of an instruction

Expliciteness characterizes the degree of accuracy of an instruction. Several marks, independently of the domain, contribute to making more explicit an instruction:
- when appropriate: existence of means or instruments,
- pronominal references as minimal as possible, and predicate argument constructions as comprehensive as possible,
- length of action explicit when appropriate (*stir for 10 minutes*),
- list of items to consider as explicit and low level as possible (*mix the flour with the sugar, eggs and oil*),
- presence of an argument, advice or warning,
- presence of some help elements like images, diagrams, etc.
- presence of elaborations, illustrations or goal specification,
- presence of a frame or a condition to limit the scope of the action.

Those criteria may be dependent on the domain, for example length of an action is very relevant in cooking, somewhat in do-it-yourself, and much less in the society domain. Similarly as for d, each item counts for 1 at the moment, expliciteness e therefore ranges from 0 to 8. The expliciteness rate is $t_i = e/8$ to keep it in [0,1]. Note also that the higher t_i is, the more chances the instruction has to succeed since it is very explicit and has a lot of details.

Now, if we consider the product $d_i \times (1 - t_i)$, the more it tends towards 1, the higher the risk is for the action to fail. Therefore, when d_i is high, it is also necessary that t_i is high to compensate the difficulty. Given that d_i remains unchanged (if the instruction cannot be simplified), the strategy is then to increase t_i as much as possible.

3. A DOMAIN-DEPENDENT ANALYSIS OF RISKS

A number of factors of risk are clearly domain-dependent. The difficulty is to be able to identify and evaluate risks without any access to a deep semantic analysis of the different actions of the domain at stake since this is seldom available.

In a first stage, as an exploration, our strategy is to extract from a large corpus of documents of the domain, for each action, the set of warnings associated with it. An action is characterized by a verb and its object argument(s), whatever their position in the instruction. Following argumentation theory, instructions with warnings have the following form: instruction because warning, as in
Carefully plug-in the mother card vertically, otherwise you will damage the connectors, where the otherwise section is the support: it indicates the risks of not doing the action correctly. In this work, if the action is 'plug-in the mother card' the risks are the list of those warnings associated with it over the whole corpus.

4. PERSPECTIVES

In this short paper, we presented the main lines of a preliminary approach to risk identification in procedures. This is a huge problem in the industry, to prevent accidents (humans and ecological). We proposed a simple solution to capture domain dependent knowledge acquired from procedure warnings. Obviously, this is just one useful facet of the problem, since a lot of knowledge is implicit and almost never expressed. Our users estimates is that we cover about 40% of the risks using this approach.

5. REFERENCES

[1] Amgoud, L., Parsons, S., Maudet, N., *Arguments, Dialogue, and Negotiation*, in: 14th European Conference on Artificial Intelligence, Berlin, 2001.

[2] Di Eugenio, B. and Webber, B.L., Pragmatic Overloading in Natural Language Instructions, International Journal of Expert Systems, 1996.

[3] Fontan, L., Saint-Dizier, P., Analyzing the explanation structure of procedural texts: dealing with Advices and Warnings, STEP conference, Venice, August 2008.

[4] Moens, M-F , Boiy, E. , Mochales Palau R. , Reed, C., *Automatic Detection of Arguments in Legal Texts*, in Proceedings of the Eleventh International Conference on Artificial Intelligence and Law, ACM Press, NY, 2007.

[5] Pollock, J.L., Knowledge and Justification, Princeton university Press, 1974.

[6] Reed, C., Generating Arguments in Natural Language, PhD dissertation, University College, London, 1998.

[7] Takechi, M., Tokunaga, T., Matsumoto, Y., Tanaka, H., *Feature Selection in Categorizing Procedural Expressions*, IRAL2003, pp.49-56, 2003.

[8] Walton, D., Reed, C., Macagno, F. (eds), *Argumentation Schemes*, Cambridge University Press, 2008.