Parsing Procedural Texts For How-to Question Answering: a principled-based approach.

Estelle Delpech, Patrick Saint Dizier

A Model for Processing Procedural Texts

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

August 3, 2007

This paper presents ongoing work dedicated to parsing the textual structure of procedural texts. We propose here a model for the intructional structure and criteria to identify its main components: titles, instructions, warnings and prerequisites. The main aim of this project, besides a contribution to text processing, is to be able to answer procedural questions (How-to? questions), where the answer is a well-formed portion of a text, not a small set of words as for factoid questions.

1. Credits

We thank the French ANR-RNTL program through the TextCoop project for supporting this research. We are grateful in particular to Claude de Loupy, Francoise Gayral, Thierry Poibeau, Frederik Cailliau and Eli Murguia for their discussions and support.

2. Introduction

The main goal of this work is to be able to answer procedural questions, which are questions whose induced response is typically a fragment, more or less large, of a procedure, i.e., a set of coherent instructions designed to reach a goal. Recent informal observations from queries to Web search engines tend to show that procedural questions is the second largest set of queries after factoid questions (de Rijke, 2005). This is confirmed by another detailed study carried out by (Yin, 2004). Procedural question-answering systems are of much interest both to the large-public via the Web, and to more technical staff, for example to query large textual databases dedicated to various types of procedures.

Answering procedural questions thus requires to be able to extract not simply a word in a text fragment, as for factoid questions, but a well-formed text structure which may be quite large. Thus, the techniques used for factoid questions do not seem adequate to deal with the problem at hand. It is clear that a different approach should be adopted. We believe that the use of text grammars is a more appropriate and a more precise manner for representing and recognizing procedural knowledge in a text.

Analysing a procedural text requires a dedicated discourse analysis, e.g. by means of a grammar. Such grammars are not very common yet due to the complex intertwinning of lexical, syntactic, semantic and pragmatic factors they require (see e.g. functional discourse grammars and systemic grammars) to get a correct analysis. Discourse grammars have basically a top-down organization, they take discourse acts as their basic units, instead of just words, they account for the structure and for the interactions between these acts and they require a relatively elaborated conceptual representation as output. Such a grammar must capture the discourse cohesion, possibly the communicative intentions, as well as the discourse organization, e.g. in terms of plans.

Procedural texts explain how to execute procedures. In our perspective, procedural texts range from apparently simple cooking recipes to large maintenance manuals (whose paper versions are measured in tons e.g. for aircraft maintenance). They also include documents as diverse as teaching texts, medical notices, social behavior recommendations, directions for use, assembly notices, do-it-yourself notices, itinerary guides, advice texts, savoir-faire guides etc. Even if procedural texts adhere more or less to a number of structural criteria, which may depend on the author's writing abilities and on traditions associated with a given domain, we observed a very large variety of realisations, which makes parsing such texts quite challenging.

Procedural texts explain how to realize a certain goal by means of actions which are at least partially temporally organized. Procedural texts can indeed be a simple, ordered list of instructions to reach a goal, but they can also be less linear, outlining different ways to realize something, with arguments, advices, conditions, hypothesis, preferences. They also often contain a number of recommendations, warnings, and comments of various sorts. The organization of a procedural text is in general made visible by means of linguistic and typographic marks. Another feature is that procedural texts tend to minimize the distance between language and action. Plans to realize a goal are made as immediate and explicit as necessary, the objective being to reduce the inferences that the user will have to make before acting. Texts are thus oriented towards action, they therefore combine instructions with icons, images, graphics, summaries, preventions, advices, etc.

Research on procedural texts was initiated by works in psychology, cognitive ergonomics, and didactics. Several facets, such as temporal and argumentative structures have then been subject to general purpose investigations in linguistics, but they need to be customized to this type of text. There is however very little work done in Computational

Linguistics circles. The present work is based on a preliminary experiment we carried out (Delpech et ali. 07), where a preliminary structure was proposed.

From a methodological point of view, our approach is based on (1) a conceptual and linguistic analysis of the notion of procedure and (2) a mainly manual corpus-based analysis, whose aim is to validate and enrich the former.

3. The structure of procedural texts

The answer to a procedural question is a well-formed fragment of a text, it includes in general a sequence of instructions linked by various markers (e.g. coordinators, temporal marks) or typographical marks (e.g. comma, dot, newline). In this section we develop an analysis of the segment title-sequence of instructions. This analysis is quite formal, but it is a necessary step before any form of processing.

In our approach, the instructional structure of procedural texts is composed of the following items:

- titles, hierarchically organized, which express a goal to reach, realized by the instructions that follow,
- instructions, associated with titles. However, instructions are not just lists of actions to perform. They often form a complex structure, presented hereafter, where there are main and subordinate instructions, comments, etc. We will therefore be here essentially concerned with an extended view of the structure of instructions, that we call **instructional compounds**,
- lists of prerequisites and warnings, besides those included into instructional compounds.

Let us essentially, in this contribution, focus on the instructional compound structure, which is, by far, the most complex element. It has a relatively well organized discourse structure, composed of several layers, which are:

- The **justification and explanation structure**, which has wider scope over the remainder of the compound, indicates motivations for doing actions that follow in the compound (e.g. *in your bedroom, you must clean regularly the curtains...*, which here motivates actions to undertake).
- The **instruction kernel structure**, which contains the main instructions. These can be organized temporally or just be sets of actions (as, for example, in social behavior texts, where instructions are rather unordered lists of advices). Actions are identified most frequently via the presence of action verbs (in relation to the domain) in the imperative form, or in the infinitive form introduced by a modal. Instructions may be subject to various conditions, and deontic and illocutionary parameters. We observed also a number of forms of subordinated instructions associated with the main instructions. These are in general organized within the compound by means of rhetorical relations, that we introduce below.
- The **deontic and illocutionary force structures**: consist of marks that operate over instructions, outling different parameters:
  - deontic: obligatory, optional, forbidden or impossible, alternates (or),
  - illocutionary and related aspects: stresses on actions: necessary, advised, recommended, to be avoided, etc.
- The **conditional structure**: introduces conditions over instructions in the compound or even over the whole instructional compound.
- The **rhetorical structure** whose goal is to enrich the kernel structure by means of a number of subordinated aspects (realized as propositions, possibly instructions) among which, most notably: causality, enablement, motivation, argument for, circumstance, elaboration, instrument, precaution, manner. The rhetorical structure is in general composed of instructions (satellites) related to the instructions in the kernel.

An instructional compound must contain at least one action verb in the imperative or a modal followed infinitive verb (in French). A few, less frequent forms have also been observed like the use of the impersonal pronoun 'on'. Several such verbs in a given sentence (e.g. coordinated) form a kernel composed of several actions which are closely related. Instructional compounds are separated by means of two main devices:

- punctuation (mainly end of sentence marks) or typodispositional marks (in .html, for example): new paragraphs, different elements in an enumeration. These are the most frequent separators.
- linguistic marks that indicate a strong break. Among these we have temporal marks that introduce a new temporal phase (glosses from French): next, after 2 hours; aspectual verbs: begin by, resume, etc.; fixed forms: this being done, you can now proceed, etc. strong breaks may also be conditional marks.

It is not necessarily easy, on the basis of linguistic marks, to make a distinction between marks that introduce a clear separation between instructional compounds from those which structure internally an instructional compound. These latter marks are weak separators or they convey an idea of continuity between instructions. Weak marks are, for example: then, finally, now, when, etc. Our strategy is to isolate two compounds when the separation mark is sufficiently strong.

The general strategy to identify instructional compounds is as follows. Any new paragraph starts an instructional compound if it contains at least one action verb in one of the forms given above (imperative or modal + infinitive). If so, the paragraph is traversed till a relevant punctuation of strong break mark is found. When this is so, the instructional compound ends and a new one is hypothetized. Sentences in a paragraph with no action verb are bound to the previous compound.

Let us now give two illustrative examples (translated from French). Here is a text extracted from the 'Home' domain: *In the bedroom, it is necessary to clean curtains.*

These are cleaned first with a vacuum-cleaner to remove dust, then, if they are in cotton, they can be washed in the washing machine at 60 degrees; if they are white, it is even recommended to add some bleech so that they look whiter. With some starch, they can be easily ironed.

The sequence: *In the bedroom, it is necessary to clean curtains* is analyzed as a justification of the actions to undertake. The next portion: *These are cleaned first with a vacuum-cleaner to remove dust, then, if they are in cotton, they can be washed in the washing machine at 60 degrees. If they are white, it is even recommended to add some bleech so that they look whiter* is the instruction kernel, where the last two instructions are associated with conditions. Finally, *With some starch, they can be easily ironed.* is an advice.

The second example introduces a subordinate instruction: *A window appears to allow you to define the properties of the server, a password is the asked, choose it with at least 6 digits. It is better to define it now since it will be asked to the user.* The first proposition describes a circumstance, it is followed by the kernel: *a password is asked*, note that this is a kind of indirect action (please provide a password), and then followed by a subordinate action of type 'precaution': *choose it with at least 6 digits*. The text ends by a motivation or justification.

4. Segmentation of instructions and titles

The segmentor has several goals:

- First, to tag terminal discourse elements: instructions, titles (viewed as the expression of goals), warnings stated outside instructions, prerequisites, and connectors.
- Via the tagging, to allow for the identification in a large text of zones which are more procedural than others (large texts may be verbose and contain non procedural elements such as comments or historical considerations), allowing then to focus the search of responses on a certain text area.

It is clear that the form of the linguistic objects and the criteria required to recognize instructions, compounds and subordinated instructions largely varies over application domains, textual genres and the targeted audiences. Our strategy was to define several sets of criteria, valid for a group of domains that share common discourse forms for describing procedures. Our approach was to proceed by 'domain aggregation'. For example, we first considered samples from coocking recipes texts, and defined the segmentation criteria. Then, we considered other domains which turn out to have a close structure: 'do it yourself' and video game solutions. At a certain stage, we get a stable set of criteria which can be implemented as an automaton. We plan to have in the end a small number of automata, each encoding the discourse structure of a group of domains.

Finally, each type of object we have to segment requires a different approach to segmentation, because the identification criteria are very different. We briefly present them below.

4.0.1. Dealing with Instructions

As far as instructions are concerned, the model is a relatively straightforward finite state automaton based on the recognition of verbs, together with their morphology (basically infinitive + modal or imperative forms are quite frequent incoocking recipes, but inflected tenses are found in other domains), and their semantic class (action verbs and subclasses) since semantic class may be used to identify different types of instructions; deverbals and predicative nouns are also relevant. Various classes of marks may also be of interest, for example to identify modalities, deontic situations or illocutionary force. Most frequent marks are: modal marks ('you must do'), reminders ('do not forget to'), performance marks ('care about doing'), marks describing optionality or advices ('it is preferable to'), injunctive forms, and adverbs of manner.

4.1. Recognizing Titles

Recognizing titles is much more challenging. They have in general the form of an instruction (e.g. mounting your computer), but with a different layout. Recognizing titles is crucial for answering questions.

Titles are first identified by the typography: bold font, possibly underlined, or via the use of dedicated html marks (h1, etc.). When this is not possible, elements such as the number of words and the proximity to an instructional zone of a certain density are good heuristics. In general, titles are much shorter than instructions. Finally, we can also rely on the level of generality of the verbs used in titles, which are more generic than those found in instructions (we use the Volem verb base for that purpose). In fact, titles can be viewed as 'super-instructions', this distinction being however highly domain dependent.

Prerequisites as well as warnings may also have titles. However, these two latter objects have a different typology (although they may also contain instructions), which allows us to make the distinction among types of titles, and to isolate those effectively governing instructions, to be interpreted as denoting goals.

A second problem is to identify the hierarchy of titles, which occurs in most texts of a certain length. Identifying such a hierarchy allows us to associate more precisely sequences of instructions to a goal. Since dedicated html tags are not so frequent to discriminate titles, we must rely on other factors, which are very delicate to handle, among which:

- presence of capital letters, or size in number of words (quite frequently 4 to 5 words, with no pronominal references),
- level of the verbs in titles: higher titles contain more generic verbs,
- identification of islands of instructions which share a quite large number of common words (entailing a certain thematic cohesion of instructions below a title, as in Centering Theory).
- identification of summaries or introductions below titles which contain words present in subtitles.

However, these criteria largely vary from one domain and author to another, and results are somewhat inconsistent. So far, we can identify 2 levels, and it may be sufficient to answer procedural questions from texts which are not too large. Gping beyond requires a much deeper analysis, and it is not clear a whether general solution can emerge.

Another kind of difficulty is that titles are often elliptic (e.g. the verb is missing). In some situation they may be just absent, therefore, we may need, to answer questions, to be able to reconstruct these, via some form of inference on the set of instructions it heads.

4.1.1. Dealing with Warnings

Warnings as well as arguments are introduced by a range of specific verbs often in the imperative form or by negative connectors, which is quite easy to identify in French. Warnings can appear at almost any position in the text. Here are a few examples:

- negative connectors: *sous peine de, sinon, car sinon, sans quoi*, etc. (otherwise, under the risk of),
- risk verbs: *risquer, causer, nuire, commettre*, etc.
- prevention verbs: *éviter, prévenir*, etc.
- negative expressions: *de facon à ne pas, pour ne pas, pour que ... ne ...pas*, etc. (in order not to).

4.1.2. Indentifying Discursive marks

Temporal marks are the most frequent marks, they include: precedence, overlap, inclusion, parallelism, etc. They are mainly realized by means of adverbs, prepositions, conjunctions, aspectual verbs and propositions describing the realization of an event. Marks are annotated by the TreeTagger and typed via a predefined list we have elaborated.

Causal marks are particularly rich and diverse. They are used to relate a goal to a set of instructions, or to specify within an instruction its aim; causal marks are also used to identify objectives, warnings and various forms of preventions, consequences and some forms of conclusions.

Besides these two main classes of marks, we noted a few conditionals and alternative marks. These are often prepositions or semantically closely related to the semantic typology specific of prepositions. To identify and interpret them, we use the PrepNet framework (www.irit.fr/recherches/ILPL/prepnet.html).

4.1.3. Global Architecture

The automaton first recognises instructions, then titles, warnings and prerequisites. Segmentation is confronted to several difficulties, among which:

The result of the segmentor is a representation based on XML tags of the following form (simplified for readability):

```
<procedure>
<title> poser une tringle a rideaux </title>
<warning>
attention a ne pas perdre des elements
       </warning>
<prereq> les regles de base .... </prereq>
<warning>
disposez de suffisemment d'espace
       </warning>
<instr compound>
1. tracer la hauteur de la tringle,
         </instr compound>
<instr compound>
2. couper la tringle a la bonne
            longueur. </instr compound>
etc.
 </procedure>
```

This representation is still quite simple and straightforward. It needs further elaborations, e.g. to deal with scoping problems which are not fully resolved. We also view this representation as a kind of dependency structure between the different constituents of the text.

At the moment, we are still updating the implementation (carried out in Perl and in Prolog) with the goal of refining the linguistic criteria necessary to accurately recognize instructional compounds, titles, warnings and prerequisites. The results we get are rather good, but they are very difficult to evaluate precisely: readers may indeed differ on their segmentation judgements for instructional compounds. Evaluating the recognition of titles is however much easier.

5. Perspectives

This short paper relates ongoing work on parsing procedural texts on various domains, with the aim of responding to procedural questions in natural language. The implementation proposed so far are preliminary and allow us to explore the various types of problems one may encounter when dealing with text grammars. The corpus considered is of a rather modest size, but with quite diverse structures. It is a development corpus, allowing us to better analyse the behavior of the different components we have developed. Outputs are checked manually at this stage: this is a quite challenging task, since most texts include more than one hundred tags, and since the boundaries of instructional compounds may be debatable. The investigations on marks is still a largely open problem since some marks may be quite pragmatic.

The global structure of a procedural text (composed of at least a main title, subtitles, prerequisites, warnings, instructional compounds) is represented by means of discourse dependency relations. In (Delpech et ali. 07), we propose a syntactic model based on elements of generative syntax for the structure of procedural texts which is fine but somewhat rigid. An approach based on dependencies, while keeping the same principles and categories allows us to have a more flexible representations, allowing some forms of gaps in the links. Similarly, we need to introduce some forms of underspecifications, in particular in relation with the difficulty of organizing titles and subtitles, in order to allow flexible forms to express scope (e.g. of a title over instructional compounds, of a condition over instructions). By default, warnings and prerequisites have scope over the text that follow them.

6. References

Aouladomar, F., Saint-Dizier, P., *An Exploration of the Diversity of Natural Argumentation in Instructional Texts*, 5th International Workshop on Computational Models of Natural Argument, IJCAI, Edinburgh, 2005.

Delin, J., Hartley, A., Paris, C., Scott, D., Vander Linden, K., *Expressing Procedural Relationships in Multilingual Instructions*, Proceedings of the Seventh International Workshop on Natural Language Generation, pp. 61-70, Maine, USA, 1994.

Delpech, E., Murguia, E., Saint-Dizier, P., *A Two-Level Strategy for Parsing Procedural Texts*, VSST07, Marrakech, October 2007.

Kosseim, L., Lapalme, G., *Choosing Rhetorical Structures to Plan Instructional Texts*, Computational Intelligence, Blackwell, Boston, 2000.

De Rijke, M., *Question Answering: What's Next?*, the Sixth International Workshop on Computational Semantics, Tilburg, 2005.

Hovy, E., Hermjakob, D., Ravichandran, D., *A Question/Answer Typology with Surface Text Patterns*, Proceedings of the DARPA Human Language Technology Conference (HLT), San Diego, 2002a.

Maybury, M., *New Directions in Question Answering*, The MIT Press, Menlo Park, 2004.

Moldovan, D., Harabagiu, S., Pasca, M., Milhacea, R., Goodrum, R., Grju, R., Rus, V., *The Structure and Performance of an Open-Domain Question Answering System*, Proceedings of the 38th Meeting of the Association for Computational Linguistics (ACL), Hong Kong, 2000.

Yin, L., *Topic Analysis and Answering Procedural Questions*, Information Technology Research Institute Technical Report Series, ITRI-04-14, University of Brighton, UK, 2004.