

# A COMPREHENSIVE SOLUTION TO PROCEDURAL KNOWLEDGE ACQUISITION USING INFORMATION EXTRACTION

Ziqi Zhang, Victoria Uren and Fabio Ciravegna  
*Department of Computer Science, University of Sheffield, Sheffield, U.K.*

**Keywords:** Information extraction, Knowledge acquisition, Procedural knowledge, Instructional text, Semantic, Ontology.

**Abstract:** Procedural knowledge is the knowledge required to perform certain tasks. It forms an important part of expertise, and is crucial for learning new tasks. This paper summarises existing work on procedural knowledge acquisition, and identifies two major challenges that remain to be solved in this field; namely, automating the acquisition process to tackle bottleneck in the formalization of procedural knowledge, and enabling machine understanding and manipulation of procedural knowledge. It is believed that recent advances in information extraction techniques can be applied compose a comprehensive solution to address these challenges. We identify specific tasks required to achieve the goal, and present detailed analyses of new research challenges and opportunities. It is expected that these analyses will interest researchers of various knowledge management tasks, particularly knowledge acquisition and capture.

## 1 INTRODUCTION

Procedural knowledge defines sequences of activities designed to achieve an objective. They are commonly embedded in the form of “instructional texts” (a term often used interchangeably with “procedural texts”) and are heavily relied upon when learning to perform new tasks to use new devices (Paris et al, 2005). Typical examples of instructional texts include cooking recipes, car maintenance guides, product usage manuals, and teaching texts.

The importance of procedural knowledge has sparked the interest of knowledge management researchers, who have dedicated considerable amounts of work to relevant research. Among these, many works have employed corpus analyses to study the topological, grammatical and rhetorical structures of instructional texts to understand what elements are essential to compose effective instructions (Kosseim, 2000; Aouladomar, 2005a; Aouladomar & Saint-Dizier, 2005; Bielsa & Donnell, 2002). Others have exploited findings from these works to develop semi-automatic Natural Language Generation (NLG) systems to help users create readable instructions (Power et al, 1998; Tam et al, 1998). While these early NLG systems adopt a knowledge elicitation process in which domain

experts are extensively involved to provide the knowledge to systems, Brassier & Linden (2002) and Paris et al. (2002) recognise the “knowledge acquisition bottleneck” of such processes and propose to use information extraction technologies to automatically extract structured procedural knowledge from largely available heterogeneous resources, including unstructured descriptive texts.

Despite the wide spectrum of relevant research, we identify several limitations of these works. Firstly, little effort has been made to establish a comprehensive set of techniques for automating the acquisition of procedural knowledge from textual data, which are commonly available existing repositories of procedural knowledge. Although some research has been carried out towards this direction (Brassier & Linden, 2002; Paris et al., 2002; Paris et al., 2005), their studies are restricted to solving sub-tasks of the entire process. Secondly, studies on instructional texts have been focusing on recognising the compositional elements of instructions (e.g., activities, objects, purposes) and their hierarchical relationships (e.g., sequences and transitions) with no link to the specific domain in question. It is believed that capturing the domain specific semantics of instructions, in addition to compositional elements, will become increasingly

important to support machine understanding and manipulation of procedural knowledge to assist users with their tasks (Sabou et al, 2009). These points are discussed further in the next section.

This paper proposes using information extraction techniques to address these issues as the next major research challenge in procedural knowledge acquisition and information extraction. The remainder of this paper is structured as follows: Section 2 discusses these issues in details and outlines research challenges; Section 3 identifies key tasks required to tackle the challenges and proposes a system architecture that integrates some information extraction techniques to achieve the goal; Section 4, 5, 6 describe these tasks in details, and identify new research challenges and opportunities involved in individual tasks. Section 7 concludes this paper.

## 2 RESEARCH CHALLENGES

Previously, the goal of most studies on procedural knowledge has been the generation of readable instructions for users, who are assumed to possess the necessary domain knowledge to understand and follow the instructions. Recent research on smart environments and smart products (Sabou et al., 2009) has set out a new challenge for procedural knowledge acquisition. It has been recognised that today the increasing complexity of products and associated tasks has hampered the usability of products. For example, modern cars are equipped with many new functionalities (e.g., audio equipment) and the instruction manual can be difficult to read (e.g., in hundreds of pages) (Mühlhäuser, 2008); likewise, cooking a meal may involve multiple kitchen appliances, each requiring different domain specific knowledge (Sabou et al., 2009) that users may not necessarily possess. To address this issue, the research advocates knowledge management technologies to equip products with the capabilities of understanding and manipulating procedural knowledge such that they are smart enough to assist users with complex tasks, to guide users through a process, and to handle exceptions and substitutions. Enabling such capability requires procedural knowledge to be defined in machine understandable forms within the domain-specific context using ontologies, which would enable knowledge fusion (Preece et al., 1999) from heterogeneous sources and automatic reasoning.

The second research challenge concerns the need for (semi-) automated means to tackle the long-recognised “knowledge acquisition bottleneck”

(Welty & Murdock, 2006). On the one hand, increasing amount of relevant information become documented and available in various sources such as internet websites (e.g., eHow.com), do-it-yourself books, user manuals (e.g., car manuals), and recipe books, enabling the average layman to learn new skills and perform complex tasks. On the other hand, to provide formalised and machine understandable procedural knowledge it is essential to automate the extraction of relevant information from these heterogeneous sources and turning them into structured forms. A comprehensive set of techniques must be applied to address different stages in the process, from extracting topic-specific instructional text to semantifying the instructions.

The interest in tackling these challenges has been brought up in research on question answering. Previously, the research has focused on responding to fact-like questions. A recent trend has however, diverted to solving procedural questions (Aouladomar, 2005; Murdock et al, 2007), and enabling machine understanding and reasoning of questions in order to provide cooperative answers (Benamara, 2004). The first aspect is closely related to the increasing demand for automatic retrieval and acquisition of instructional information from heterogeneous sources; the second addresses the capability to reason and manipulate knowledge by incorporation of semantic technologies (e.g., ontologies) to provide alternative answers when definitive answers are not available (i.e., being cooperative).

We encourage researchers to develop new methods or adapt existing techniques to address these challenges. Particularly, it is believed that recent advances in information extraction have delivered a suite of techniques that may contribute to a comprehensive solution. In the following, a set of key tasks that are required to build a complete system for procedural knowledge acquisition is proposed. They are analysed to identify relevant techniques and the new challenges and opportunities facing each individual technique. We believe this will facilitate researchers to choose their research topics of interest.

## 3 TOWARDS A SOLUTION

We propose a system architecture that takes textual sources (e.g., a recipe book) of instructional information and domain ontologies as input, and that produces structured procedural knowledge individually identified by the task objectives, and semantically bound to the domain of interest (e.g.,

cooking or car maintenance). We acknowledge the existence of procedural knowledge in other media such as pictures and videos; however, we aim to focus on the textual resources because of their richer availability and complementary role to other media.

Three major sub-tasks that are involved in the processing are identified, including 1) extracting passages of instructional information from original documents (*passage extraction*); 2) recognising instructional components and their hierarchical relations (*procedural analysis*); 3) domain-specific annotations of instructions (*semantic tagging*). The process is illustrated in Figure 1 below. Sections 4 to 6 in the following discuss each sub-task in detail.

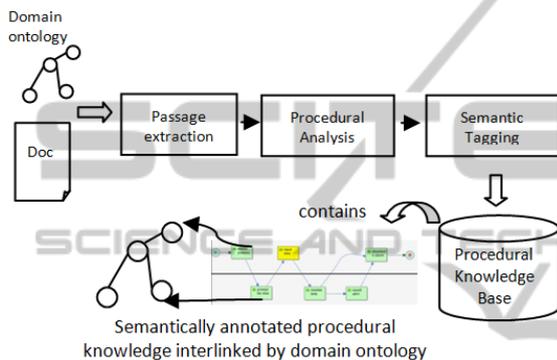


Figure 1: Illustration of the envisioned process of procedural knowledge acquisition.

## 4 PASSAGE EXTRACTION

Given an input document containing instructional information, the first step is to identify passages that describe particular instructions for individual tasks. Previous research on instructional texts (Brasser & Linden, 2002; Paris et al., 2002; Paris et al., 2005) has all assumed the availability of such passages. However, passage identification and extraction is a non-trivial issue in procedural knowledge acquisition, and must be addressed as the first step in practical scenarios. For instance, a car manual contains instructions about carrying out different tasks; it also contains a good proportion of non-instructional data, such as precautions, warnings, illustrations, introductions. Likewise, a recipe book contains procedures of making different meals, each of which may contain non-instructional sections describing ingredients or background information about the recipe. Obviously, procedural knowledge acquisition from such data must firstly identify the boundaries between passages and extract passages of interest only.

Passage identification and extraction is a topic that has been actively studied in the information retrieval (Wang & Si, 2008) and question answering communities (Tiedemann, 2007). The main goals are defined as segmenting documents into potential passages and extracting relevant ones that satisfy a retrieval need. Traditionally, the segmentation strategy has focused on using physical document structure such as paragraphs, and sentences; defining passages of fixed length; and detecting topic shifts between passages (Oh et al., 2007). The extraction of passages is then formulated as a retrieval task, in which segmented passages are ranked according to their similarity (semantic or distributional) to a query or question (Tiedemann, 2007).

Essentially, the first step in procedural knowledge acquisition can be formulated as an information retrieval problem that asks “find all passages that are instructional texts in the document”, such that the classic passage segmentation and extraction techniques can be applied. However, the scope of the query is more general than topic-specific queries (e.g., “find the texts about cooking sea food pizza”) that current research focuses on. It is in fact so general that traditional content-based feature modelling methods such as using content words and their meanings may fail. To illustrate, consider the sample instructions shown in Figure 2a and 2b, both of which are extracted instructions from a *Ford 2006 Focus* car manual (from <http://www.focusplanet.com/>).

To adjust your mirrors:

1. Rotate the control clockwise to adjust the right mirror and rotate the control counterclockwise to adjust the left mirror.
2. Move the control in the direction you wish to tilt the mirror.
3. Return to the center position to lock mirrors in place.

Figure 2a: Sample instruction of “Adjust mirrors”.

**Climate Controls**

To aid in side window defogging/demisting in cold weather:

1. Select .
2. Adjust the temperature control to maintain comfort.
3. Set the fan speed to the highest setting.
4. Direct the outer instrument panel vents toward the side windows.

To increase airflow to the outer instrument panel vents, close the vents located in the middle of the instrument panel.

Figure 2b: Sample instruction of “Climate control”.

Both samples are good candidate instructional texts and therefore, should be extracted as valid passages of interest for *procedural analysis*. It is obvious that the two passages are about different topics. The similarity of the two passages cannot be captured by modelling their content. Therefore, to effectively extract both passages and other similar ones, alternative features must be explored.

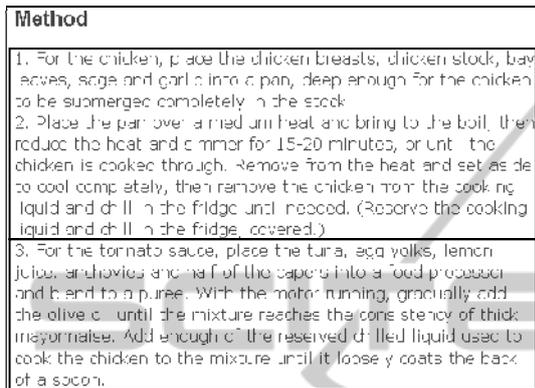


Figure 3: Sample instruction of a recipe extracted from BBC.co.uk.

Consider again an instruction of a recipe shown in Figure 3. Although it is evident that the whole piece of text is a valid passage describing a coherent topic of cooking a meal, it is also worth noting that sub-steps 1-2 and 3 are also valid passages describing sub-procedures of the task. Being able to identify such nested structure within instructional texts enables better understanding of a procedure. Therefore, passage segmentation and extraction should also cope with such nested relations.

Our initial analyses of a collection of online recipes and car manuals reveal that although content-based features used in classic passage extraction may be of limited use, alternative cues may be exploited to capture the regularities in such data. For example, instructional texts are often characterised by sequences of activities, identified by bullet point structures. Within a restricted domain such as a single document or a set of similar documents, such texts are often formatted by a distinctive set of styles (e.g., font family, font size, list structure), or identified by headings of distinctive styles. Also, previous studies on instructional texts have shown that the language used in composing instructions often employs a “stereotypical” set of grammatical, sentential and rhetorical structures (Kosseim, 2000; Aouladomar, 2005). For example, Linden & Brassler (2002) argue

that the phrasal structure “to do something” often indicates the goal to be achieved by a procedure.

These cues bear structural information of a passage that are potentially effective at identifying the presence of procedural information and their boundaries, and may complement or even replace the content-based features. We believe exploiting such features for passage extraction can lead to promising results in procedural knowledge acquisition.

## 5 PROCEDURAL ANALYSIS

Once proper passages of instructions have been extracted, the next step in procedural knowledge acquisition is the identification of compositional elements in a procedure, such as actions, objects, and actors; and their hierarchical relations, such as sequence, and transitions. Multiple terminologies have been used to describe the whole or parts of this analysis process, including “hierarchical decomposition” (Tam et al., 1998), “rhetorical structure analysis” (Kosseim, 1998; Aouladomar, 2005a), “task concept” and “task relations” (Paris et al., 2002). For the sake of simplicity we refer to this process as “*procedural analysis*”.

The majority of previous works on instructional texts address this process. Studies have shown that instructions are often composed of a limited set of building blocks, such as objectives, participants, objects and actions, sequences and transitions. And the language used for describing instructions exhibits regular grammatical, sentential and rhetorical structures. The phenomenon is referred to as “stereotypical content and structure” by Kosseim (2000), and is also found common across different languages (Bielsa, 2002; Kosseim, 2000). Therefore, intuitively, one could apply Natural Language Processing (NLP) techniques to perform structural analysis, such as tokenisation, part of speech tagging, sentence parsing and discourse analysis. Next, a limited set of rules exploiting such regularities could be built to perform procedural analysis. Typically, Brassler & Linden (2002) and the extension of their work by Paris et al. (2005) built a finite-state grammar that essentially employs rules to identify key elements in a procedural description and their relations. All of these systems isolate domain-specific knowledge from the analysis, by focusing on recognising “closed” classes of words such as verbs and connectives as indicators of key elements or relations of a procedure. For example, verbs most likely indicate an action, and words that follow the verbs are likely to be objects.

Although such methods have domain independence, they suffer from two limitations. Firstly, the analysis is not at semantic level but is literal and syntactic. They may recognise a verb as an action and the following noun as an object, they do not analyse the domain specific meaning of the verb and the object, and thus do not enable understanding of the procedure. This issue will be discussed in detail in Section 6. Secondly, ignoring domain-specific knowledge may cause skewed performance of systems. For example, in biomedical information extraction, a specifically trained part of speech tagger must be used for biomedical texts in order to capture special term and grammar compositions in the domain (Tsuruoka et al., 2005). Domain specific terms (e.g., “On” and “Off” represent control switches on an MP3 player; “tbsp” means “table spoon”, and “oz”, “gram” are weight units in recipe texts) are found common in domain-specific instructions. Without domain-specific knowledge it can be difficult to recognise these terms and, therefore, corresponding objects and actions. On the other hand, if certain forms of domain-specific knowledge become available, one can use them to guide the recognition of compositional elements and their relations in *procedural analysis*.

Automatically constructing and extracting domain-specific knowledge from heterogeneous sources has been a constant focus in the research of information extraction. Many techniques such as lexicon construction (Ando, 2004), term recognition (Ananiadou, 1994) and entity recognition (Cimiano & Völker, 2005) are designed for this purpose. It is believed that incorporating these techniques can aid the task of *procedural analysis*.

## 6 SEMANTIC TAGGING

Previous works on procedural knowledge acquisition terminate after the *procedural analysis* stage, since the focus has been generating human readable instructions. As argued in the previous sections, it is envisioned that the future studies on procedural knowledge acquisition should address machine understanding and manipulation of knowledge in order to provide systems with the capability of guiding users through complex tasks to achieve their objectives. The benefits of such capability could be illustrated using several short scenarios. For example, a *VW Golf* car user may query a procedural knowledge base for “how to adjust side mirror”. Assuming the knowledge base contains only instructions of a 2006 *Ford Focus* model,

understanding that *Golf* and *Focus* are both models of cars but belong to different manufacturers, the knowledge base can suggest the similar procedure (such as Figure 2a) found in the *Ford* car knowledge base as a potentially valid substitute. In a cooking scenario, understanding that “tbsp”, “oz”, “gram” are all weight units, the knowledge base can convert weights between different measures and adjust interaction with users depending on their preferences, such that one is no longer troubled with finding out how much is a “tbsp” or “oz”. In a smart products environment, participating devices in a procedure may collaborate to proactively take over certain sub-tasks from users. For example, in helping users with the roast turkey recipe, knowing that “pre-heat oven to 200 degree for 30 mins” and “steam vegetables using steamer for 20mins” are two sub-processes each involving one domain specific “smart” product participant (oven and steamer), the knowledge base may delegate the sub-processes to those devices, which are capable of performing sub-procedures automatically without requiring user intervention (e.g., the oven automatically starts, sets temperature, and triggers timer to monitor the process).

Although enabling such high level of intelligence may be long term research that requires integration from many scientific disciplines, advances of relevant technologies have already enabled researchers to take the first step (Mühlhäuser, 2008). Procedural knowledge acquisition must take one step further from *procedural analysis* (Section 6) by binding the knowledge acquired to domain specific semantics. Essentially, domain-specific ontologies must be used to index procedural knowledge, such that knowledge fusion and reasoning capabilities can be supported. The process depends on a number of information extraction tasks, such as domain specific entity recognition (Cimiano & Völker, 2005) and ontology population (Cimiano, 2006). Although classic approaches already exist, efforts should be made to minimise a system’s dependence on human supervision (e.g., providing large amounts of examples) and addressing domain portability (i.e., extracting domain-specific knowledge from different domains). Additionally, the characteristics of instructional texts described in Section 5 may serve as useful cues in developing extraction systems.

## 7 CONCLUSIONS

This paper discusses automating procedural knowledge acquisition using information extraction techniques. The importance of procedural

knowledge has long been recognised and relevant research has been performed. However several limitations of these works have been identified, namely, lack of comprehensive set of (semi-) automatic techniques to tackle the acquisition bottleneck, and the capability of enabling machine understanding and manipulation of procedural knowledge. It is believed that these are the two pressing issues that must be addressed in future research on procedural knowledge acquisition, the evidence for which can be found in the recent trends of relevant research and the emergence of smart product research. This paper argues for applying various information extraction techniques to address different stages in procedural knowledge acquisition to form a comprehensive solution. Specific tasks required to achieve the goal have been identified with new challenges and opportunities analysed. It is believed that these will create new interest in the research of information extraction and procedural knowledge acquisition, also potentially other aspects of knowledge management, such as knowledge representation.

## ACKNOWLEDGEMENTS

Part of this research has been funded under the EC 7th Framework Programme, in the context of the SmartProducts project (231204).

## REFERENCES

- Ando, R., (2004). Semantic Lexicon Construction Learning from Unlabeled Data via Spectral Analysis. *Proceedings of CoNLL'04*.
- Aouladomar, F., (2005). A Preliminary Analysis of the Discursive and Rhetorical Structure of Procedural Texts. In *Symposium on the Exploration and Modelling of Meaning*
- Aouladomar, F., Saint-Dizier, P., (2005). An Exploration of the Diversity of Natural Argumentation in Instructional Texts. In *Proceedings of IJCAI'05 Workshop on Computational Models of Natural Argument*. p.69-72
- Benamara, F., (2004). Cooperative question answering in restricted domains: the WEBCOOP experiment. *The ACL Workshop on QA in Restricted Domains*.
- Bielsa, S., Donnell, M., (2002). Semantic Functions in Instructional Texts: A Comparison between English and Spanish. In *Proceedings of the 2nd International Contrastive Linguistics Conference*, p.723-732
- Brasser, M., Linden, K., (2002). Automatically Eliciting Task Models From Written Task Narratives. In *Proceedings of the 4<sup>th</sup> International Conference on Computer-Aided Design of User Interfaces*, p.83-90
- Cimiano, P., (2006). Ontology learning and population from text: algorithms, evaluation and applications, Springer.
- Cimiano, P., Völker, J., (2005). Towards large-scale, open-domain and ontology-based named entity classification, *Proceedings of RANLP'05*.
- Ananiadou, S. (1994) A methodology for automatic term recognition. *Proceedings of COLING '94*.
- Kosseim, L., (2000). Choosing Rhetorical Structures to Plan Instructional Texts. In *Journal of Computational Intelligence*, Vol. 16, p408-455
- Max Mühlhäuser, (2008) Smart Products: An Introduction. In: *Constructing Ambient Intelligence - Aml 2007 Workshops*, pp. 158-164, Springer Verlag.
- Murdock, V., Kelly, D., Croft, W., Belkin, N., Yuan, X., (2007). *Identifying and improving retrieval for procedural questions*. In *Information Processing & Management*, Vol. 43 (1), pp. 181-203
- Oh, H., Myaeng, S., Jang, M., (2007). Semantic passage segmentation based on sentence topics for question answering. *Journal of Information Sciences*, Vol. 177
- Paris, C., Colineau, N., Lu, S. (2005) Automatically Generating Effective Online Help. In *International Journal on E-Learning*. Association for the Advancement of Computing in Education
- Paris, C., Linden, K., Lu, S. (2002). Automated Knowledge Acquisition for Instructional Text Generation. In *Proceedings of SIGDOC'02*.
- Power, R., Scott, D., Evans, R. (1998). What You See Is What You Meant: direct knowledge editing with natural language feedback. In *ECAI'98*.
- Preece, A., Hui, K., Gray, A., Marti, P., Bench-Capon, T, Jeans, D., Cui, Z. (1999). The KRAFT architecture for knowledge fusion and transformation. In *Expert Systems*. Springer
- Sabou, M., Kantorovitch, J., Nikolov, A., Tokmakoff, A., Zhou, X., and Motta, E., (2009). Position Paper on Realizing Smart Products: Challenges for Semantic Web Technologies, In: *The 2nd International Workshop on Semantic Sensor Networks*, collocated with ISWC'09
- Tam, R., Mulsby, D., Puerta, A. (1998). U-TEL: A Tool for Eliciting User Task Models from Domain Experts. *Proceedings IUI'98*, ACM
- Tiedemann, J., (2007). Comparing document segmentation strategies for passage retrieval in question answering. In *Proceedings of RANLP 07*.
- Tsuruoka, Y., Tateishi, Y., Kim, J., Ohta, T., McNaught, J., Ananiadou, S., Tsujii, J. (2005) Developing a Robust Part-of-Speech Tagger for Biomedical Text, *Advances in Informatics, 10<sup>th</sup> Panhellenic Conference on Informatics*
- Wang, M., Si, L., (2008). Discriminative probabilistic models for passage based retrieval. *Proceedings of the 31st annual international ACM SIGIR*
- Welty, C., Murdock, J., (2006). Towards Knowledge Acquisition from Information Extraction. *ISWC2006*.