DOI: 10.15439/2014F161
ACSIS, Vol. 2

Knowledge Extraction from professional e-mails

Nada Matta, Hassan Atifi, François Rauscher
ICD/Tech-CICO, Université de Technologie de Troyes
12 rue Marie Curie, BP. 2060, 10010 Troyes Cedex, France
nada.matta@utt.fr, Hassan.atifi@utt.fr, Fracois.Rauscher@utt.fr

***Abstract*—Some professional e-mails contain knowledge about how actor face problem in order to realize projects. This type of knowledge is produced in cooperative activity. Representing project knowledge leads to structure link between coordination, cooperative decision-making and communication. The main objective of our work is to extract knowledge from daily work. So the main questions of our research are:**

- **Can we extract knowledge from professional e-mails?**
- **If so, which type of knowledge can be represented?**
- **How to link this knowledge to project memory?**

We present in this paper our first work in this aim. Our hypothesis is tested on a software development application.

***Index Terms*—Knowledge Engineering, Knowledge Management, Project memory, Traceability, Professional e-mails, Pragmatics analysis.**

I. Introduction

CURRENTLY, Designers use knowledge learned from past projects in order to deal with new ones. They reuse design rationale memory to face new problems. Knowledge Management provides techniques to enhance learning from the past [5]. Their approaches aim at making explicit the problem solving process in an organization. Their techniques are inherited mainly from knowledge engineering. So, we find in these approaches in one hand, models representing tasks, manipulated concepts and problem solving strategies, and in the other hand, methods to extract and represent knowledge. We note for instance MASK [7], [14] and REX [11] methods. These methods are used mainly to extract expertise knowledge and allow defining profession memories.

But, design projects involve several actors from different fields. These actors produce knowledge when interacting together and take collaborative decisions. So, it is important to also tackle this type in knowledge, which is generally volatile.

We deal, in our approach with this type of knowledge, called Project memory [13]. Project memory must represent organizational and cooperative dimension of knowledge. Current techniques used in Knowledge management, based on expert interviews are not adapted to extract these dimensions of knowledge. To tackle knowledge produced in collaborative activity, we need techniques that help to extract knowledge from daily work. In this paper, we present a technique that help to extract knowledge from professional e-mails. The presented approach allows structuring extracted concepts and linking them to the project context. We use pragmatics analysis and knowledge engineering techniques for this aim.

II. Project Memory

A project memory is generally described as "the history of a project and the experience gained during the realization of a project" [13]. It must consider mainly (Fig. 1.):

- The project organization: different participants, their competences, their organization in sub-teams, the tasks, which are assigned to each participant, etc.
- The reference frames (rules, methods, laws, ...) used in the various stages of the project.
- The realization of the project: the potential problem solving, the evaluation of the solutions as well as the management of the incidents met.
- The decision making process: the negotiation strategy, which guides the making of the decisions as well as the results of the decisions.

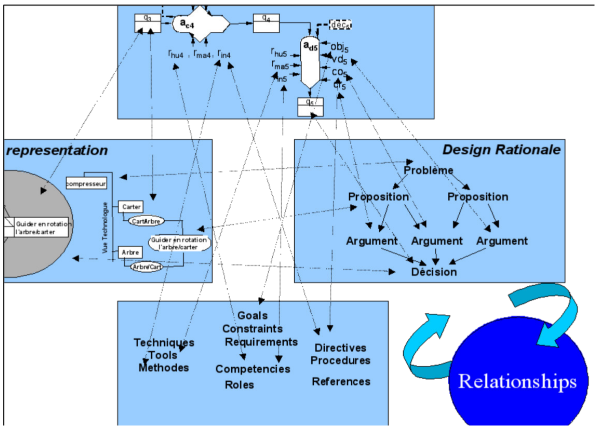


Fig. 1. Project memory

Often, there are interdependence relations among the various elements of a project memory. Through the analysis of these relations, it is possible to make explicit and relevance of the knowledge used in the realization of the project. The traceability of this type of memory can be guided by design rationale studies and by knowledge engineering techniques.

978-83-60810-58-3/$25.00 © 2014, IEEE

III. Pragmatics analysis

The act of request has been extensively studied in the field of theoretical linguistics (Searle 1969), intercultural and inter-language pragmatics [2], NLP community on automated speech act identification in emails [3], [11] etc. However, as pointed by *Rachele De Felice et al. [4]* there is very little work concerned with data other than spoken language and few researches seem to fully respond to requirements of being sufficiently general, non-domain specific, and easily related to traditional speech acts. In addition, few researchers have focused their research requests in business written discourse (workplace email communication). Lambert et al 2010 try to create tools that assist email users to identify and manage requests contained in incoming and outgoing email. Atifi et al. [1] analyze email effectiveness from the professional's point of view by mixing two kinds of analysis: a content analysis of interviews of professionals and a pragmatic and conversational analysis of emails. *Rachele De Felice et al. [4]* propose a global classification scheme for annotating speech acts in a business email corpus based on traditional speech act theory described by Austin and Searle [15].

.

IV. Related work on E-mails analysis

Several approaches study how to analyze e-mails as a specific discourse. We note for instance, tagging work [17], in which Yelati presents techniques that help to identify topics in e-mails, or the use of zoning segmentation in [10]. Other works use natural language processing in order to identify messages concerning tasks and commitment [8]. They parse verbs and sentences in order to identify tasks and they track messages between senders and receivers.

Even there is lot of work on pragmatics, which study dialogue and distinguish techniques in order to identify speech intention (Patient/doctor dialogue analysis [8]), coding dialogue scheme [Core et al, 1997], etc. Pragmatics analysis of e-mails uses only some of these methods like ngrams analysis by Carvalho in [16], Verbal Response Mode scheme by Lampert in [10]or a custom coding scheme like De Felice [4].

Techniques studying e-mails, often do not consider the context of discussions, which is important to identify speech intention. We deal with our work with professional e-mails, extracting from projects. So, we mix pragmatics analysis and topic parsing and we link this type of analysis to project context (skill and role of messages senders and receivers, project phases, and deliverables, etc.) in order to keep track of speech intention. As pragmatics analysis shows, there is not only one grid to analyze different types of speech intention. In project memory, we look for problem solving, design rationale, coordination, etc. In this study, we focus on problem solving and we build an analysis grid for this purpose.

V. Project knowledge extraction from e-mails

The main objective of our work is to extract knowledge from daily work. So the main questions of our research are:

- Can we extract knowledge from professional e-mails?
- If so, which type of knowledge can be represented?
- How to link this knowledge to project memory?

To answer these questions, we analyze professional e-mails related to projects. In last studies, we identify a structure to analysis coordination messages [12]. Based on pragmatics analysis, we defined a grid to structure coordination messages based on the main act to do (inform, request, describe, etc.) and the objects of coordination (task, role, product, etc.). In this paper, we will go ahead and define an approach that helps to extract knowledge from professional e-mails. So, we identify firstly step by step how to isolate important messages and how to analyze them. Knowledge from e-mails, as knowledge produced in daily work, cannot be very structured. It is related closely to context. In our work, we focus on knowledge produced during project realization. We will show in our method how information from project organization help in e-mails knowledge extraction.

A. Classification of e-mails

Firstly, we have to identify important messages (Fig. 2). For that, we have to gather messages in subjects. Then, we can identify the volume of messages related to each subject. Then we analyze only messages that heave more then 4 answers; we believe that knowledge can be extracted based on interaction. Finally, we link the messages to be analyzed to project phases.

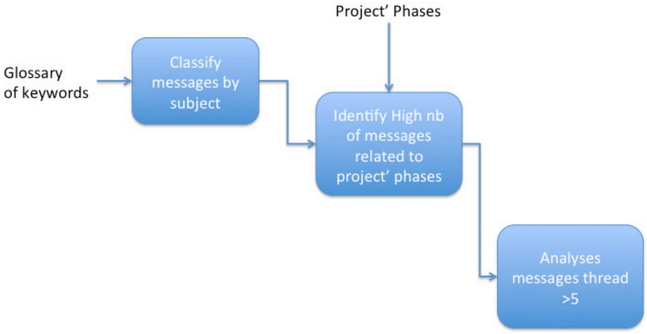

Project' Phases
Glossary of keywords
Classify messages by subject
Identify High nb of messages related to project' phases
Analyses messages thread >5

Fig. 2. First e-mails analysis

B. Messages analysis

For each message thread (message and answers), we identify (Fig. 3) :

- Information to be linked to organization:
  - Authors, To whom, In Copy
- Information about phases:
  - Date and hour of messages and answers
- Information about product:
  - Topic and joined files
- Information about message intention:
  - Main speech act and intention of message

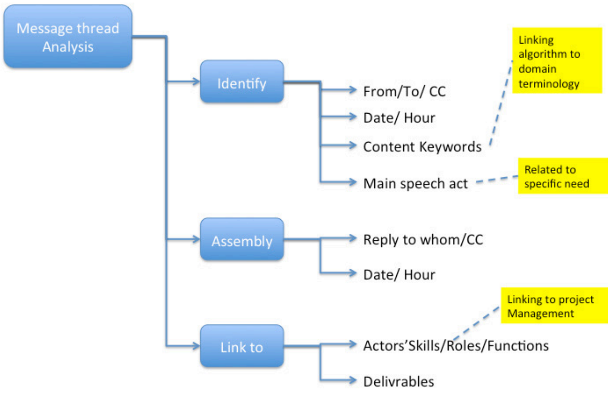


Fig. 3. Analysis of messages

By linking messages to project organization, we help in making sense of interactions between actors. In fact, the role and skill of messages' senders and receivers help to analyze the role of the message in problem solving and the nature of the content (solution answering a problem, proposition discussions, coordination messages, etc.). In the same way, linking messages to phases help to identify main problems to deal in each phase of the same type of projects.

As first work, we focus our speech act analysis on problem solving by identifying request and solution. So, we identify first speech acts that help to localize a request in a message (**Erreur ! Source du renvoi introuvable.**). Then, we study the organization of related messages thread in order to identify the solution proposed (if it exists) to the request. Our analysis is based first on pragmatics in order to characterize request speech act, and that by identifying request verbs and forms. In the present study we limited our research to the analysis of the act of requesting in problem solving sequences.

From a pragmatic point of view, a request is a directive speech act whose purpose is to get the hearer to do something in circumstances in which it is not obvious that he/she will perform the action in the normal course of events [15]. By introducing a request, the speaker believes that the hearer is able to perform an action. Request strategies are divided into two types according to the level of interpretation (on the part of the hearer) needed to understand the utterance as a request. The two types of requests include direct request and indirect requests. The request can be emphasized either projecting to: 1- the speaker (Can I do X?) or 2- the hearer (Can you do X?). A direct request may be use an imperative, a performativity, obligations and want or need statements.
An Indirect request may use query questions about ability, willingness, and capacity etc. of the hearer to do the action or use statements about the willingness (desire) of the speaker to see the hearer doing x. At last, for us, a grammatical utterance corresponds to only one speech act as in TABLE1.

TABLE 1.
GRID OF REQUEST SPEECH ACT

Request Form	**Linguistic form**	**Examples**
Direct request	Imperative	Do x
	Performative	I am asking you to do x.
	Want or Need statements.	I need/want you to do x
	Obligation statements	You have to do x
Indirect request	Query questions about ability of the hearer to do X	Can you do x? Could you do x?
	Query questions about Willingness of the Hearer to do X	Would you like to do x?
	Statements about the willingness (desire) of the speaker	I would like if you ca do X I would appreciate if you can do X

Then, we complete our analysis by from one side identifying answers verbs and from another side, linking answers to actors' role and skills and also joining files. The date of answers can be an indicator of several elements in the organizations: engagement, difficulty of time spending of solution, stress and multi-responsibilities, etc. We aim at analyzing in the future the frequency of answers.

VI. Example

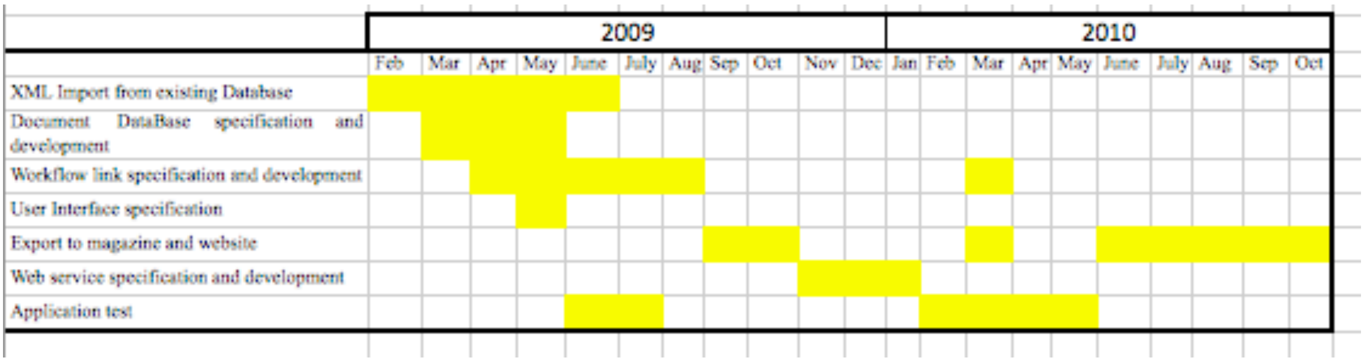
A. Example description

INFOPRO Business Publishing Company asked a software Company to develop a workflow tool that helps journalists to edit their articles and to follow the modification of the journal. The period of the project was more than one year. Nearly all negotiations and discussions were through e-mails. In this project, the actors were:

- SRA: an editing responsible (skill: law and management, Role: Contractor)
- JBJ: Information System Manager (Skill: Information system, Role: Contractor)
- FX: Information System Developer (Skill: Software Engineering, Role: Development manager)
- CV: Prototyping (Skill: Human Machine Interface, Role: User Interface Modeling)
- RT: Information System Developer (Skill: Software Engineering, Role: Sub-contractor)

Principles phases of the project were (TABLE 2):

TABLE 2.
PHASES OF THE PROJECT



	2009											2010									
	Feb	Mar	Apr	May	June	July	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	June	July	Aug	Sep	Oct
XML Import from existing Database	🟨	🟨	🟨	🟨	🟨																
Document DataBase specification and development		🟨	🟨	🟨																	
Workflow link specification and development			🟨	🟨	🟨	🟨	🟨							🟨							
User Interface specification				🟨																	
Export to magazine and website								🟨	🟨					🟨			🟨	🟨	🟨	🟨	🟨
Web service specification and development										🟨	🟨	🟨									
Application test					🟨	🟨							🟨	🟨	🟨	🟨					

B. E-mails Analysis

As first step, we identify messages topics based on e-mails subjects. In our project, we identify main discussions topics based on keywords:

- XML : structuration, tag, tree, xsd, dtd, schema
- BDD : Data base, table, editing part, code part
- Interface workflow : UI, Workflow, User Interface, login, user management,
- Code : Insurance Code, Legifrance, auto code, vehicle code, mutu code, chapter, article
- Document : new collection, construction, document
- Export paper : Indesign, layout, mapping, tag indd, indd template
- Export site: export web, web tag, web format, dtd web
- Export Author: word, author, xslt author
- Services: update legi, Legifrance update, FTP
- Word : macro word, addin, web service, word 2007, wordlink

Based on these topics we use a Lucene based algorithm, to compute distance and similarity between words in order to identify main topics of messages (boosting email subject importance compared to email body (Fig. 4). It is to be noted that some preprocessing occurred before on raw email body to remove duplicated answers):

$$\text{score}(q,d) = \text{coord-factor}(q,d) \cdot \text{query-boost}(q) \cdot \frac{V(q) \cdot V(d)}{|V(q)|} \cdot \text{doc-len-norm}(d) \cdot \text{doc-boost}(d)$$

Fig. 5 shows first step of analysis of these messages; in which we show senders and receivers and their skills, topics of messages and date of messages.

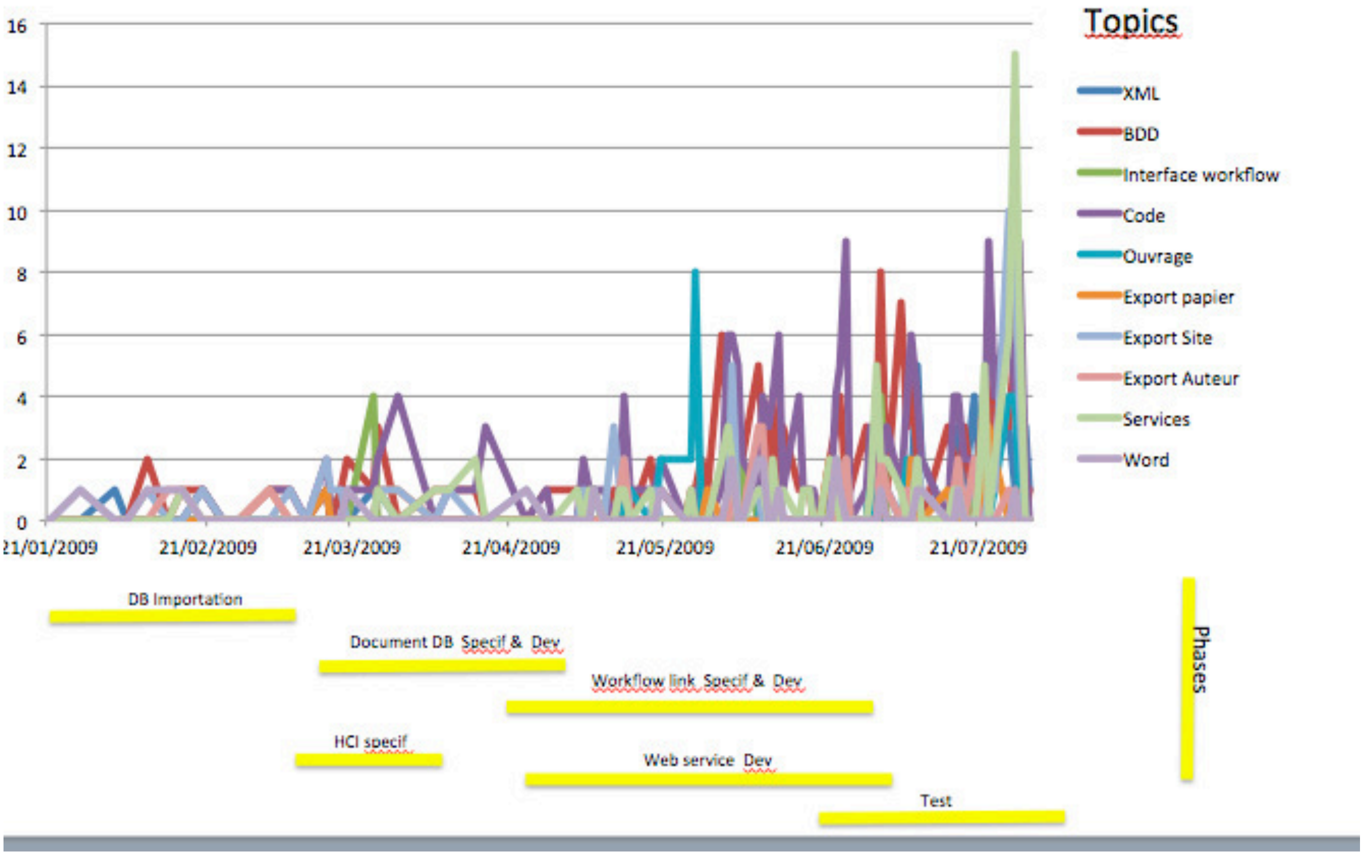
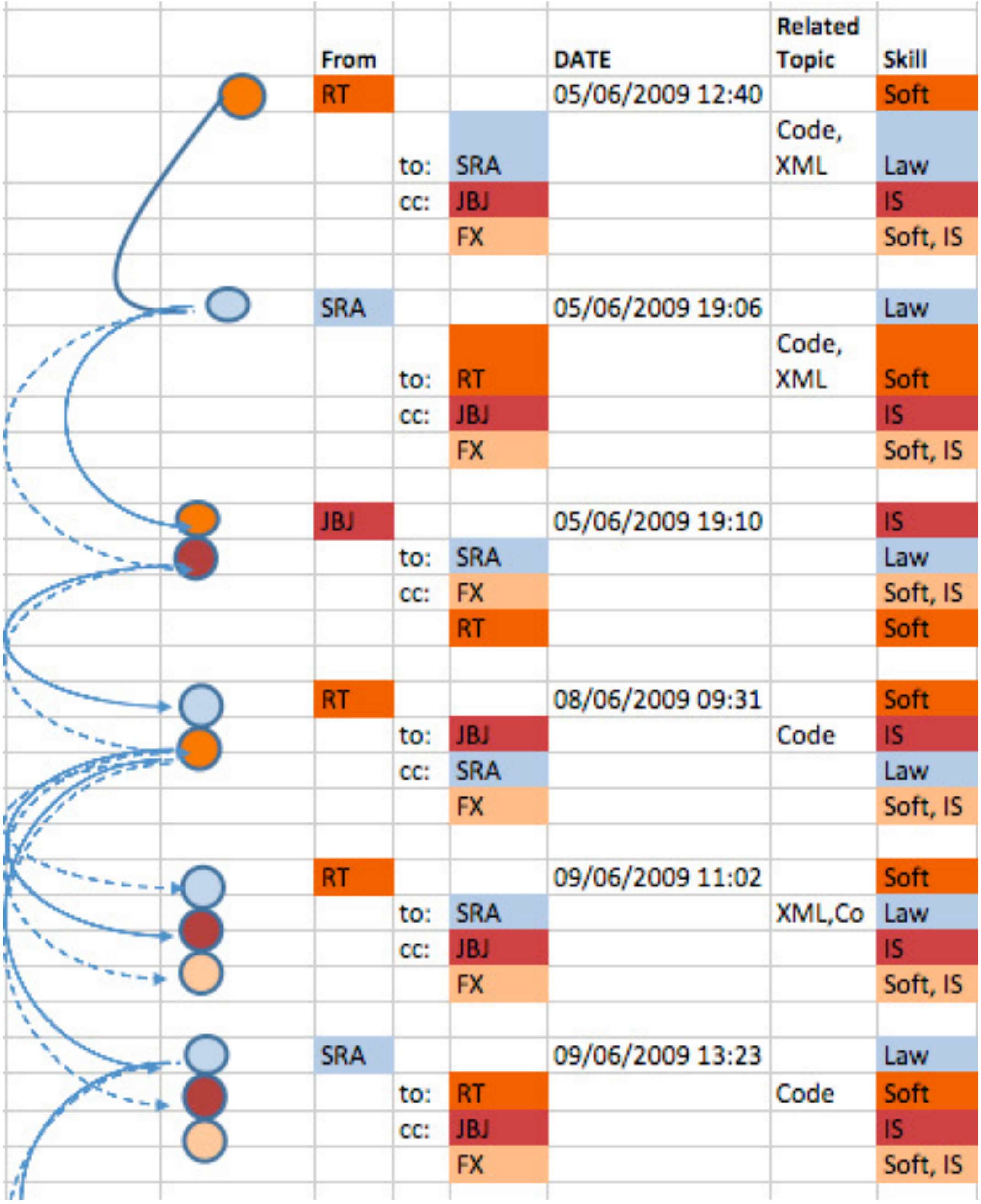


Fig. 4. Topics analysis



From			DATE	Related Topic	Skill
RT			05/06/2009 12:40		Soft
	to:	SRA		Code, XML	Law
	cc:	JBJ			IS
		FX			Soft, IS
SRA			05/06/2009 19:06		Law
	to:	RT		Code, XML	Soft
	cc:	JBJ			IS
		FX			Soft, IS
JBJ			05/06/2009 19:10		IS
	to:	SRA			Law
	cc:	FX			Soft, IS
		RT			Soft
RT			08/06/2009 09:31		Soft
	to:	JBJ		Code	IS
	cc:	SRA			Law
		FX			Soft, IS
RT			09/06/2009 11:02		Soft
	to:	SRA		XML,Co	Law
	cc:	JBJ			IS
		FX			Soft, IS
SRA			09/06/2009 13:23		Law
	to:	RT		Code	Soft
	cc:	JBJ			IS
		FX			Soft, IS

Fig. 5. First analysis of messages: representing of senders/Recievers/Copy, date and actors role and skill

To analyze messages text, we use pragmatics in order to identify problem and solution discussions. For that, we identify Request messages based on Request speech acts. Then, we identify related answers messages. In these messages we look for senders skill and joined files. So, we identify for the “Annexes” topic, in which there are 23 messages, related topics are XML and code. Messages were during 12 days from 5th until 17th June. They concern workflow development phase. Based on the Request-Answer grid and role actors, we analyze messages, in order to identify problem-solving intentions. So, we identify for instance, the problem Insurance Text extraction. SRA; the editing responsible (contractor) asks FX to extract Insurance text in a good format. When FX; the Information System Developer (Development manager) answer him, we suppose that as an answer, based on the role of sender of message and the main topic. We consider also joined files as part of this answer. Fig. 6 shows this example.

From			**Date**	**Sentence elements**	**Related Topic**	**Function**
SRA			2009-06-05 12:40:46.	I put in "Bold", what **I need**:		Request
	to:	FX		1- *Issurances*		
	cc:	JBJ		2- Text without tags Texte in XML files	Code	
		CV		3- Tag Pb : Text outside tag in XML	XML, Code	
		RT		4- Tag Pb <b> is opened and not closed, as same as, tag is badly imbricated		
FX			2009-06-05 19:06:34.			Answer
	to:	SRA		1- *Issurances*		
	cc:	JBJ		I propose to convert: Xpress format in XML	XML	
		CV		Beware, the text will contain a lot of error blanc, "enter" and image	Code	
		RT		I can transform it on enriched XML	XML	
				It containts a lot of reference, so we have to compose with links		

Fig. 6. Example of messages analysis

VII. Conclusion

The aim of our study is to identify knowledge from daily work. In this paper, we show that it is possible to study professional e-mails for this aim. We consider e-mails as specific discourse. So we use pragmatics generally used to analyze discourse and to categorize it to identify knowledge from professional e-mails. Our hypothesis is can we identify a grid as guide to analyze professional e-mails? If so, can the result be relevant as project knowledge?

Based on this hypothesis, we know that pragmatics intention must be based to context. So, we consider the project context from different aspect: organization and environment. We believe that this context is very helpful to clarify ambiguity of sentence analysis. We show in the example how sender/receiver role can identify problem-solving answer. Adding this analysis to the identification of keywords of messages, as topics can be a first step, towards a structuring of knowledge: Problem related to a topic, possible answers.
We will continue to validate this work on other type of projects. This work can open to identify other grid analysis like: engagement of actors, design-rationale, coordination [12], etc.

Finally, this study is a part of our work on project memory: Keeping track and structuring knowledge in daily work realization of project. We developed techniques to extract knowledge from project meetings [6] and to identify occurrences in order to identify concepts in project memory.

References

[1] Atifi H., Gauducheau N., Marcoccia M.2011.The Effectiveness of Professional Emails: Representations and Communicative Practices , in proceedings of 13th Conference of the International Association for Dialogue Analysis, Dialogue and Representation, Montréal.

[2] Blum Kulka, Shoshana, Juliane House, and Gabriele Kasper (eds.) 1989. Cross-cultural pragmatics: Requests and apologies. Norwood : Ablex Publishing.

[3] Carvalho, Vitor and William Cohen. 2005. On the collective classification of email “speech acts”. Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 345–352. New York: Association for Computing Machinery.

[4] De Felice, R; Darby, J; Fisher, A; Peplow, D; (2013) A classification scheme for annotating speech acts in a business email corpus. ICAME Journal , 37 71 – 105

[5] Dieng R. , Corby O., Giboin A. and Ribière M., Methods and Tools for Corporate Knowledge Management, in Proc. of KAW'98, Banff, Canada. 1998.

[6] Ducellier G., Matta N., Charlot Y., Tribouillois F., “Traceability and structuring of cooperative Knowledge in design using PLM”, Knowledge Management and collaboration Special Issue of International Journal of Knowledge Management Research and Practices, Vol.11, No.1, 2013, pp 53-61.

[7] Ermine J.L., La gestion de connaissances, J.-L. Ermine.- Hermès sciences publications, 2002.

[8] Kalia K.A., Identifying Business Tasks and Commitments from Email and Chat Conversations, tech. report, HP Labs, 2013

[9] Lampert, Andrew, Robert Dale and Cecile Paris. 2010. Detecting emails containing requests for action. Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL), 984–992. Association for Computational Linguistics.

[10] Lampert, A., Dale, R., & Paris, C. (2006). Classifying speech acts using Verbal Response Modes. In Proceedings of the 2006 Australasian Language Technology Workshop (ALTW2006), 34-31

[11] Malvache P., Prieur P., Mastering Corporate Experience with the REX Method, Proceedings of ISMICK’93,

International Synopsium on Management of industrial and corporate knowledge, Compiegne, October, 1993.

[12] Matta N., Atifi H., Sediri M. Sagdal M. , Analysis of interactions on coordination for design projects, IEEE proceedings of the 5th International Conference on Signal-Image Technology and Internet based Systems, Kula Lumpur, December, 2010.

[13] Matta, N., Ribière, M., Corby, O., Lewkowicz, M., et Zacklad, M. Project Memory in Design, Industrial Knowledge Management - A Micro Level Approach. Springer-Verlag : Rajkumar Roy, 2000.

[14] Matta N., Ermine J-L., Gérard Aubertin, Jean-Yves Trivin, Knowledge Capitalization with a knowledge engineering approach: the MASK method, In Knowledge Management and Organizational Memories, Dieng-Kuntz R., Matta N.(Eds.), Kluwer Academic Publishers, 2002.

[15] Searle, J.R. (1969). Speech Acts. An Essay in the Philosophy of Language. Cambridge: Cambridge University Press.

[16] Vitor R. Carvalho and William W. Cohen. 2006. Improving "email speech acts" analysis via n-gram selection. In *Proceedings of the HLT-NAACL 2006 Workshop on Analyzing Conversations in Text and Speech* (ACTS '09). Association for Computational Linguistics, Stroudsburg, PA, USA, 35-41

[17] Yelati, S.; Sangal, R., "Novel Approach for Tagging of Discourse Segments in Help-Desk E-Mails," Web Intelligence and Intelligent Agent Technology (WI-IAT), 2011 IEEE/WIC/ACM International Conference on , vol.3, no., pp.369,372, 22-27 Aug. 2011