Relation Extraction in an Intelligence Context

Bénédicte Goujon

Thales Research & Technology - RD 128
91767 Palaiseau Cedex - France
benedicte.goujon@thalesgroup.com

Abstract

Our aim is to produce structured information from unstructured texts. To do so, we want to automatically extract explicit relations between entities from texts. Our work is constrained by the targeted intelligence domain, where users have no expertise in linguistics and cannot work with linguists for confidentiality reasons. We have developed a first prototype called Sem+ which extracts binary relations between entities from texts. It was mainly used on French corpus, but can be used for English. It was developed in a first time to automatically supply knowledge base. Relations are extracted thanks to patterns that are defined by the end user. For example, *"Henri Konan Bédié a reçu Alassane Dramane Ouattara.*" (Henri Konan Bédié has received Alassane Dramane Ouattara) is a pattern which produces the following relation: CONTACT(Henri Konan Bédié, Alassane Dramane Ouattara). Sem+ uses a learning algorithm, based on the Hearst algorithm, to ease the pattern acquisition. Several evaluations were provided on sale and purchasing relations between companies, and on an Ivory Coast corpus. The good precision and the efficiency of the learning algorithm were motivating to improve the tool. First improvement concerns the verbal pattern management. We add general linguistic knowledge to enhance the number of relations extracted with each pattern. Also we have worked to improve the entity management, in order to identify not only proper names but also nominal expressions related to entities ("*le président ivoiren*" as well as "Laurent Gbagbo"). This work was focused on people category. Now Sem+ is being integrated into platforms. The first one is a decision support platform, where Sem+ extracts relations from texts in order to identify events. The aim of the platform is to send an alarm when several events occur. The objective is to prevent a crisis, and the current study is based on the Ivory Coast crisis of September 2002. Sem+ will also be integrated in a semantic web platform, which contains complementary tools to annotate documents and manage ontologies.
Index Terms: relation extraction, text mining

1. Introduction

The objective of this work is to produce structured information from unstructured texts. To do so, we want to automatically extract structured relations between entities using a method which will be efficient on various domains. The relation extraction task is very hard and has to be restricted to few cases. As we can observe in the last ACE evaluations [1], only one participant proposes the relation extraction on English in 2007. In our approach, we first have worked on sales and purchases relations between companies. We have also studied a corpus dealing with the Ivory Coast situation, containing various named entities (Person, Organization, …) and relations (Location, Contact, …). A first prototype called Sem+ has been developed from these studies on those two different themes.

In this paper, we first describe the Sem+ tool together with its learning algorithm, and a first evaluation. We then present how we have added general linguistic knowledge to improve this tool and the resulting relations. We also present the platforms where Sem+ is used as a relation extraction component for various information treatment needs.

2. Sem+ : our Relation Extraction tool

We present here the most important aspects of our tool Sem+, which was detailed in [2].

2.1. Objectives and context

The principle aim is to analyze unstructured texts to provide to our clients structured information that follows their specific needs. To do so, we want to extract relations between entities from texts.

Thales has developed the Idéliance tool, which is a knowledge management system based on the concept of semantic networks [3]. The conceptors wanted to enable an easy use of the tool, for users without specific notions in knowledge representation. The manipulated knowledge has the format of a triplet "subject / verb / complement", as in "Peter / is from the category / Person", "Peter / is going to / Paris", etc. The main limitation of this tool concerns the knowledge capture. For now, it must be done manually. A great improvement of this tool would be the automation of the capture of all the relations. To do so, we have worked on automatic information extraction.

An important constraint is the specificity of the users: as Idéliance is used in the intelligence domain, the user wants a tool that allows to work alone, without any specific linguistic knowledge. We cannot imagine the intervention of a linguist on sensitive data.

Our objective was to propose a demonstrator allowing an easy acquisition of the relation patterns. Relation patterns are used to extract occurrences of relations from texts. The end user is a domain expert and not a linguist. For the pattern acquisition, we wanted to use an existing learning algorithm, and adapted it if necessary to ease the task. We also wanted to test whether only specific knowledge is sufficient to extract specific information so that the first demonstrator will not manage generic linguistic knowledge.

The work described here was done on French data, but the system is also useful for English. We wanted to obtain a tool providing the less noise, so we have preferred ignoring relations than proposing bad ones. And, as our tool was developed to be plugged with Idéliance, we have only considered as structured information the relations between two entities for the moment. For example, it includes a relation of sales or purchasing between two companies, or a relation of location between a person and a place. Such

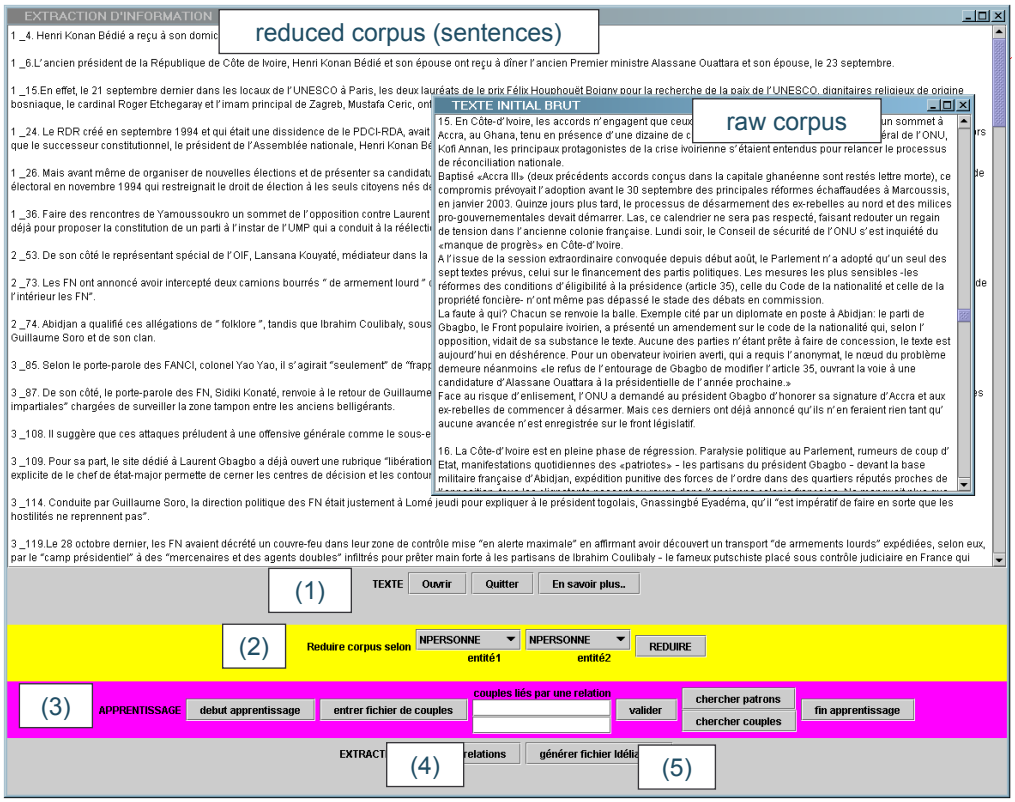

Figure 1: Sem+ Interface.

relations and the related entities are different from those managed in the ACE program [4]. In this campaign, a lot of entity types and subtypes are defined. First, we have worked with Named Entities (Person, Location) which represent a subset of the ACE entity types. Relations are also different from the ACE relations: we have specifically studied explicit relations expressed by verbs or nouns. For us, such relations are very relevant, as they are given by the writer of the text to We have studied two kinds of solutions: frameworks and ACE oriented approaches, in order to compare with our specific need.

T-REX (Trainable Relation Extraction framework) is a general software framework for supervised extraction of entities and relations from text [5]. This framework was designed so as to provide the degree of flexibility required by automatic semantic markup tasks for the Semantic Web. It has been developed as a testbed for experimenting with several extraction algorithms and several extraction scenarios, especially extraction from the web. This framework, like the Gate infrastructure [6], is not an information extraction system, but ease the building of such system. As we were used to manipulate a French linguistic environment (Intex), we have chose it for the first prototype.

SIE (Simple Information Extraction) is an information extraction system based on a supervised machine learning technique for extracting implicit relations from documents [7]. Another extraction method is presented by Zhao and Grishman [8]. It uses a kernel SVM. Those solutions learn off-line a set of data models from a specified labelled corpus, as it is defined in the ACE campaign. Such approaches are not efficient in our context as we can't have pre-tagged corpus dealing with the intelligence domain.

inform on an unusual fact or event. We didn't take into account implicit relations, expressed by word co-occurrences.

2.2. Relation extraction solutions

Several solutions are presented to solve the information extraction, and sometimes the relation extraction from texts.

2.3. Our learning algorithm

We first choose a learning method to build the linguistic patterns, rather than a declarative method and rather than a mathematical approach. The limit of the declarative methods in our context is that linguistic knowledge must be defined by a linguist after a corpus study. With a mathematical approach, the resulting co-occurrences relations are not accurate enough. For example, a relation between two persons can have a lot of meanings (CONTACT, KILL).

After the study of the three following learning systems: Rapier [9], EXDISCO [10] and Prométhée [11], we opt for the development of a specific algorithm. Our algorithm is based on the Hearst algorithm [12] but adapted to the intelligence domain as previously explained. Here are the steps of this algorithm:

1. Selection by the user of a couple of entity categories concerned by the relevant relation;
2. Capture by the user of couples of entities verifying the relation;
3. Automatic recovery of sentences containing these couples, with patterns that potentially describe the relation;
4. Selection by the user of the sentence extracts expressing the relation, capture of the relation information and the automatic transformation of the

extracts into patterns with Intex, a linguistic development environment [13].

5. Use of the patterns: recovery of new couples. Back to the step 3.

In order to simplify the reading of the corpus by the user and the extraction of the couples, the corpus is previously reduced according to the couple of entity categories that is concerned by the relation. At step 4, the user may transform the sentence, if some words are not meaningful, to express the relation by using "**" instead of the optional words.

To illustrate our approach, here is an example. We suppose that the user knows that Henri Konan Bédié and Alassane Dramane Ouattara were in contact recently (2). From this couple, the user obtains the following sentence: *"Henri Konan Bédié a reçu à son domicile parisien Alassane Dramane Ouattara."*[1] (3). The user builds then the pattern "*Henri Konan Bédié a reçu ** Alassane Dramane Ouattara*", and captures additional information: "*Henri Konan Bédié*" is the agent, "*CONTACT*" is the relation expressed by the pattern and "*Alassane Dramane Ouattara*" is the patient (4). Applied on another corpus containing "*Laurent Gbagbo a reçu hier ... Henri Konan Bédié*", this algorithm automatically produces the new relation "*Laurent Gbagbo – CONTACT – Henri Konan Bédié*". From this new couple (5), the system may identify a new sentence (3) that contains potentially another way to express the contact relation (4).

2.4. Details of the Sem+ demonstrator

We have developed the Sem+ demonstrator in order to evaluate the real interest of our algorithm for the intelligence domain. The first Sem+ demonstrator has been developed by Julia Frigière[2]. It is based on Intex [12] and is developed in Java and Perl.

Before the use of Sem+, the user has to construct its own domain-specific dictionaries, which are used to identify the entities. Here is an sample of the user dictionary containing persons:

Gbago,Laurent Gbagbo.NPERSONNE
Henri Konan Bédié,.NPERSONNE
Lansana Kouyaté,.NPERSONNE
Laurent Gbagbo,.NPERSONNE

Here are the steps for using Sem+ (see Figure 1 above).

1. Selection of a corpus, and automatic tagging of the named entities described in the dictionaries (for example, Person.dic contains "Lansana Kouyaté", Location.dic contains "Bouaké").
2. Reduction of the corpus according to a couple of entity categories (Person and Location in our example).
3. Learning step (the two sub-steps below could be applied constantly):
   - Capture of the entity couples to initiate the learning approach / recovery of new couples obtained from the captured patterns.
   - Capture of patterns associated with couples. The capture is easy, and consists of a copy-paste of the sentence extract (pattern) containing the relation, as: "*Lansana Kouyaté s'est rendu à Bouaké*"[3]. The user validates the agent, the patient, and the category of the relation expressed in the extract. For each extract, a generic transducer is automatically created with Intex to obtain for example MOVING(Person, Location) from "Person *s'est rendu à* Location".
4. Use of the pattern set to extract relations on new corpora: "*Thabo Mbeki s'est rendu à Bouaké*" => MOVING(Thabo Mbeki, Bouaké).
5. Production of a file using the Ideliance format, in order to export these relations into the knowledge management system.

2.5. Sem+ evaluations

First evaluations have been done on a financial corpus (Sale and Purchasing relations between Companies), to test the patterns coverage on a new corpus and the efficiency of the learning cyclical algorithm. On a small new corpus, using 45 patterns obtained from the learning corpus, Sem+ has identified 11 relations among 42 relations: the recall was 26% and the precision was 100%. Several identified reasons have caused the low recall. Firstly, only the sentences containing two entities are taken into account, even if sometimes relations may be built on two sentences (with anaphora). Also, as there were no linguistic analysis, the verbal expressions of the pattern were fixed, so "*X a acheté Y*" and "*X achète Y*"[4] are two different patterns. For us it is a correct result, as we want to favour the precision, but we'll have to improve it.

To evaluate the efficiency of the learning algorithm, by using 20 couples that were manually acquired (by reading of the first dispatches), we obtain automatically the capture of 32 patterns. This result is motivating, as it illustrates the efficiency of our method to easily learn new patterns.

Another evaluation has been provided on an Ivory Coast corpus, in order to validate the usability of such an approach by a user without specific linguistic knowledge. This evaluation has been done at the Land Headquarter (S.T.A.T. unit), by the Officer Cadet Cytermann[5]. He has used Sem+ on a corpus composed of journalistic articles related to the Ivory Coast. For this evaluation, no specific criterion have been used to quantify the results. After some technical adjustment and a first test, it has been concluded that Sem+ is easy to use, and that it is efficient in saving time when coupled with the Ideliance system. On this corpus, the learning algorithm was not efficient because of the small size of the acquisition corpus (120 sentences). Also, there were a lot of relations to identify (Meeting between two persons, Appointment between two persons, Moving between a person and a location, Accusation between two organisms, or between persons, etc.). The number of cases for each relation was not large enough for an efficient learning approach. This focuses on the characteristics of the initial corpus, that is essential in a learning approach: if the corpus is not large enough, it may not provide enough relation patterns to automate the information extraction.

---

[1] Henri Konan Bédié has received at his Parisian address Alassane Dramane Ouattara.

[2] *Frigière J. (2004), Information extraction by learning method, Internal report*.

[3] Lansana Kouyaté has gone to Bouaké.

[4] X has bought Y, X buys Y

[5] *Cytermann F. (2005), Évaluation du logiciel Sem+ dans le domaine du renseignement militaire, Training STAT report.*

3. Integration of general linguistic knowledge

As it was shown in the previous evaluations, Sem+ has to be improved in order to extract more relations. As we didn't use general linguistic knowledge in the first version, we have decided to add some to improve our results. We present here an algorithm that was developed to improve the verbal patterns management. We also detailed the focus on the entity management that is done via the anaphora resolving. We then present a future work that deals with the realization degree of each relation occurrences. Those improvements will be evaluated soon.

3.1. Verbal patterns management

During the previous evaluation done by a end-user, we have noticed that the pattern acquisition was a hard task. The use of no other knowledge than the user dictionaries is not sufficient to generate the various verbal expressions in relation with each verb. We propose to improve this by pre-defining the most common verbal expressions in a graph: conjugated verb; "to have", "to be able" or "to go" followed by the infinitive form of the verb; "to have" followed by the past participle form of the verb, eventually with adverbs for each case. We have developed our own code rather than using an existing parser, in order to control everything in Sem+. Here is our algorithm for the enrichment of verbal patterns, after the validation of an extract associated to a relation:

- Identification of the verbal sequence present in the extract;
- Identification of the lemma associated to this verbal sequence;
- Combination of this verb and the generic pre-defined graph containing the most common verbal expressions, and production of the complete verbal pattern;
- Production of the relation pattern, combining most of the variability of the verbal forms and the other words of the extract.

From the English example "X had received Y", we are able to automatically identify all those patterns : "X had received Y": "X is going to receive Y", "X has just received Y", "X will soon receive Y". If a verb is not in the general dictionary, the program just keep the initial extract. Two generic pre-defined graphs are used: one for the active form, and one for the passive form.

This verbal pattern management code has been implemented for French, by using the Delaf dictionary, and for English.

Our system manages patterns containing the verb between entities, as in previous examples, and patterns containing the verb after the entities (as in "*Laurent Gbagbo et Henri Konan Bédié se sont rencontrés*"[1]). Such patterns can contains two optional sequences at the most. We also have added the patterns containing a noun placed before the entities in relation, as is "*La rencontre entre Laurent Gbagbo et Henri Konan Bédié*"[2]. To evaluate this work, we have observed that previously we had to define five patterns for the "*rencontrer*" verb from the acquisition corpus. Now one pattern is enough. On the evaluation corpus, this pattern extracts six relations, which shows its efficiency.

3.2. Entity management

We have focused in a first time our work on the automatic identification of relations, using dictionaries containing proper names to identify the entities. But, this entity identification is not efficient. For example, in our corpus several expressions are used to talk about "*Laurent Gbagbo*": "*Le président Laurent Gbagbo*", "*Le président ivoirien Laurent Gbagbo*", "*Le président*", *"il"*[3], … All those expressions must be associated to Laurent Gbagbo in order to extract most of the relations, as in "*Il a rencontré Henri Konan Bédié*"[4]. In a first time, we have worked on anaphoric nominal expressions related to people. Our aim was to find a simple way to solve the identification of the referent of such expression without needed descriptions by the user. From our corpus, we have observed that the context of proper names was containing a lot of information. For example, if we have in a text "*le président ivoirien Laurent Gbagbo*", we can deduce that the expression "*le président ivoirien*" or "*le président*" can refer to Laurent Gbagbo in specific cases. An algorithm was developed to manage those anaphora[5]. In a first step, a corpus analysis extracts all the expressions preceding proper names and associates their content information (role, country, organization, …) to the corresponding entities. In a second step, a text is analysed and for each expression referring to a person without proper name ("*le président ivoirien*"), the algorithm propose a best candidate, which is referred in the preceding text and which is associated to the same information. This algorithm is currently evaluate.

3.3. Relation realization degree management

During our evaluation with potential users, we have observed that relations extracted from texts were not exact every times, as sometimes the relation has not occurred. We have identified then that an important clue in a relation extraction is the degree of realization of the relation: if we extract: CONTACT(Henri Konan Bédié, Alassane Dramane Ouattara) from the sentence "*Henri Konan Bédié devrait recevoir Alassane Dramane Ouattara*"[6], the user may consider as real a relation that has not occurred yet. Such variations on the meaning are very important in sensitive domain as intelligence or command context. A lot of relations are expressed using conditional tenses or words in order to express the uncertainty of the situation.

We want to propose the use of a specific feature associated to each relation that will express this realization degree. This will allow the distinction between the relations that have occurred (according to the text) using the keyword "certain", that may occur ("not certain"), and that will occur ("not realized"). To do so, we will exploit several linguistic markers: the verb tense (conditional means "not certain", future means "not realized"), the verbal structure ("to go"

---

[1] Laurent Gbagbo and Henri Konan Bédié have met each others

[2] The encounter between Laurent Gbagbo and Henri Konan Bédié

[3] The president Laurent Gbagbo, the president of the Ivory Coast Laurent Gbagbo, the president, he

[4] He has met Henri Konan Bédié

[5] *« Amélioration de la gestion des entités nommées pour l'extraction de relations sémantiques », Elzbieta Gryglicka, training report, 2006.*

[6] Henri Konan Bédié should receive Alassane Dramane Ouattara

followed by an infinitive means “not realized”) and the lexical markers (“may” or “to suppose” means “not certain”).

4. Sem+ integration in plateforms

In parallel with its own improvements, Sem+ is integrated into two platforms. The first integration was done previously for decision support. The second integration is in process and is related to the Semantic Web. We present here the two platforms.

4.1. Integration of Sem+ in an information management platform

The aim of a Command Support System (CSS) is to support decision by providing the Commander an edge over knowledge. In that context, we have built a global platform described in [14] that supports operators in their information analysis task. This platform contains three steps:

1. first, information is extracted from various media (video, text, audio, various signals from various sensors …);
2. second, the information is fused when necessary (if two pieces of information are dealing with a same event);
3. third, an alert is sent when necessary (if the new detected event is similar to important previous events).

Sem+ is integrated in the first phase of this platform. The approach used fo the fusion task is described in [15].

We have began the study of a corpus dealing with the Ivory Coast crisis of September 2002. The main event was on the night of September 18-19, 2002, when as many as 800 disgruntled soldiers took up arms against their country in mutiny. Our corpus contains about 15 000 texts from newspaper or press agency. From this corpus, we would like to identify some precursory events that were at the origin of the crisis, in order to provide the sending of an alarm before such a crisis. To do so, Sem+ will be useful to extract relations between people and/or organizations. By adding date and location information we would like to extract necessary information to describe each event.

4.2. Integration of Sem+ in a Semantic Web platform

The objective of the WebContent French project [16] is to propose a platform for the content management. This platform will be used to automatically discover, understand and structure the information whatever its format, and particularly the documents (XML and HTML), the Web Services and the forms. It contains various components to semantically annotate documents. Several partners have developed components for various tasks dealing with document annotation or knowledge management: document segmentation, lemmatization, parsing, filtering, classification, ontology enrichment, etc.

In this context, Sem+ will be used for the entity extraction and for the relation extraction. To do so, we want to produce relations that will be automatically used to enrich ontology. Also, we would like to include the anaphora resolution algorithm to annotate all the textual occurrences of persons. And the addition of realization degree will be helpful to distinguish events that are not described as “realized”.

This platform will be used by partners for several applications in relation with technological. For Thales, we will use this platform for our study on the Ivory Coast crisis. We will use several services complementary to Sem+ to manage the texts and the knowledge extracted from them.

5. Conclusion

We have presented the first version of our prototype called Sem+ that extract entity relations from texts. It is based on a learning algorithm that helps user without any linguistic knowledge to manage the relations between entities. We have detailed the evaluation of this version. We also have presented some improvements that were provided for the verbal management and for the entity management. Our Sem+ tool is now used into a CSS platform, and will be added to a semantic web platform. We would like to add realization degree information to the relations, and we will have to manage precisely date and location entities, as we are working on person entities to solve anaphoric expressions, in order to extract complete relations.

6. References

[1] ACE 2007 Automatic Content Extraction Evaluation Offical Results (ACE07) http://www.nist.gov/speech/tests/ace/ace07/doc/ace07_eval_official_results_20070402.htm

[2] Goujon B., Frigière J. 2005. Extraction of Relations between Entities from Texts by Learning Methods, in IST-055 Specialists Meeting on "Information Fusion for Command Support", Netherlands.

[3] Rohmer J. 2002. Représentation, Fusion et Analyse d'informations mises sous forme de réseaux sémantiques: vers le "calcul littéraire" ?, Revue REE*, N°7 Juillet 2002.*

[4] ACE 2005 Evaluation Plan, http://www.itl.nist.gov/iad/894.01/tests/ace/

[5] Iria J., Ciravegna F. 2005. Relation Extraction for Mining the Semantic Web, Dagstuhl Seminar on Machine Learning for the Semantic Web.

[6] Cunningham H., Maynard D., Bontcheva K., Tablan V. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02). Philadelphia, July 2002.

[7] Giuliano C., Lavelli A., Romano L. Simple Information Extraction (SIE), Technical report.

[8] Zhao S., Grishman R. 2005. Extracting Relations with Integrated Information Using Kernel Methods, Proceedings of the 43rd Annual Meeting of the ACL, pages 419–426, Ann Arbor, June 2005.

[9] Califf M. E., Mooney R. J. 2003. Bottom-Up Relational Learning of Pattern Matching Rules for Information Extraction, in Journal of Machine Learning Research 4*, pp177-210.*

[10] Yangarber R., Grishman R., Tapanainen P., Huttunen S. 2000. Automatic Acquisition of Domain Knowledge for Information Extraction, in Proceedings of the 18th International Conference on Computational Linguistics (COLING 2000)*, Saarbrücken, Allemagne.*

[11] Morin E. 1999. Using Lexico-Syntactic Patterns to Extract Semantic Relations between terms from Technical Corpus, TKE'99*, Innsbruck, Austria, August 99, pp. 268-278.*

[12] Hearst M. A. 1992. Automatic Acquisition of Hyponyms from Large Text Corpora, in 14TH International

Conference on Computational Linguistics (COLING 1992), *pp. 539-545.*

[13] Silberztein M., INTEX : http://msh.univ-fcomte.fr/intex/

[14] Laudy C, Mattioli J., Museux N., 2005. Cognitive Situation Awareness for Information Superiority, IST-055 Specialists Meeting on "Information Fusion for Command Support", Netherlands.

[15] Laudy C., Ganascia J.-G., Sedogbo C., High-level Fusion based on Conceptual Graphs, Fusion 2007.

[16] WebContent project description (in English) : http://www.webcontent.fr