Extraction of Relations between Entities from Texts by Learning Methods

Bénédicte Goujon and Julia Frigière
THALES Research & Technology France
RD 128, F-91767 Palaiseau Cedex
FRANCE

Benedicte.goujon@thalesgroup.com, frigierejulia@yahoo.fr

ABSTRACT

The aim of this work is to automatically extract structured information from unstructured texts, permitting their fusion in an intelligence application. In Thales, we have a knowledge management system (Idéliance) that permits us to manage entities and relations between them, but at present the user must manually capture this information. To automate such an extraction, we propose the use of a learning algorithm that we have developed after the study of the existing information extraction methods. We present the Sem+ tool that implements the algorithm, and the evaluation of this tool carried out by us and by the Land Headquarter (S.T.A.T. unit).

1 INTRODUCTION

The aim of this work is to automatically extract structured information from unstructured texts, permitting their fusion in an intelligence application. For us, a structured information is composed of a relation between two entities, for example a relation of sales or purchasing between two companies, or a relation of location between a person and a place.

In Thales, we have a knowledge management system (Idéliance) that permits us to manage entities and relations between them, but at present the user must manually capture this information. Our aim is to ease the user task by allowing the automatic acquisition of such knowledge from texts.

To automate such an extraction, we propose the use of learning methods. We briefly present our work below: we have studied existing methods and chosen the most efficient, we have developed the Sem+ system, which automates the learning method to extract information, and we have performed two evaluations on different corpus to validate our approach.

2 INFORMATION EXTRACTION BEFORE FUSION

The most important prerequisite for fusing information is to identify and extract the relevant information. In most cases this step is done manually. In this paper we describe a method to automatically extract relevant information from knowledge provided by the user.

2.1 Ideliance

In Thales we have the Ideliance tool, which is a knowledge management system based on the concept of semantic networks (Rohmer, 2002). The aim of the conceptors was to offer the easy use of such a tool, for

Goujon, B.; Frigière, J. (2006) Extraction of Relations between Entities from Texts by Learning Methods. In *Information Fusion for Command Support* (pp. 9-1 – 9-8). Meeting Proceedings RTO-MP-IST-055, Paper 9. Neuilly-sur-Seine, France: RTO. Available from: http://www.rto.nato.int/abstracts.asp.

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 working for / Thales”, etc.

The main limitation of this tool is the capture of the knowledge. At present, it must be done manually. A great improvement to this tool would be the automation of the capture of all the relations. To do so, we have worked on the automatic information extraction.

2.2 Information extraction: state of the art

Various methods are proposed to automate information extraction in various contexts. We have identified three main approaches: declarative approaches, statistical approaches and supervised learning approaches. Here is the presentation of these approaches, with comments on their interests and limits.

2.2.1 Declarative approaches

A recent work (Bouhafs-Hafsia) was done on the description of all necessary French linguistic knowledge to extract information in an intelligence context. The aim of this approach was to describe beforehand all the knowledge that is necessary to identify precise information. This description of linguistic knowledge is done by a specialist in the area from the study of a corpus.

Eleven relations were defined: Negociation, CoLocation (two persons in a same location), Confrontation, Communication, etc. Each of these relations is associated to a set of words explaining the relation :“marchander”, “traiter”, “négocier” for the Negociation; “retrouver”, “recevoir”, “contacter”, “rejoindre”, “voir”, etc. for the CoLocation, …[1]. Most of these words are verbs. These words are grouped together and associated to specific rules according to contexts that may be encountered: we can have “X et Y ont négocié …”, “X a traité avec Y …” for the Negociation verbs, “X a contacté Y”[2] for the CoLocation verbs. The implementation of these rules is based on the contextual exploration principle (Desclés, 1997).

The strong point of such an approach is the efficiency of the precise rules to capture most of the information associated with the defined relations.

Numerous weak points are associated with such a method. First, this method is expensive: each relation must be defined for each language, by a linguist after a specific study of a representative corpus. Second, if end users are concerned by another relation that was not previously defined, they can’t extract anything, they just have to wait for the linguist to provide the complete description (words, rules) of the new relation. And thirdly, this approach may be unusable in a strategic intelligence domain where end users don’t want anyone to observe and manipulate their confidential data.

2.2.2 Statistical approaches

The opposite approach is the statistical ones, which does not need any linguistic knowledge. The principle is to apply statistical calculations to identify words that are frequently together in texts, which signifies that a relation exists between these words, according to the distributional hypothesis of Harris (1971).

Such an approach has the advantage of not needing predefined specific knowledge to be usable. Also, the identified information is obtained without prejudice.

[1] to haggle over, to deal, to negociate, … to meet again, to welcome, to contact, to rejoin, to see

[2] “X and Y have negociated …”, “X has dealt with Y …” …, “X has contacted Y”

Unfortunately, relations that are obtained are sub-specified: if such an approach identifies a relation between two persons, which is the precise relation: are they friends or enemies? were they in the same place? are they members of a common group? With such methods, users have to analyse resulting data in order to identify the specific relation. Finally, it does not provide a precise automatic relation extraction.

2.2.3 Supervised learning approaches

Other approaches are based on learning methods to extract information. Most of these methods are used to ease the terminologists or knowledge engineers tasks (Aussenac-Gilles and al, 2000), the construction of dictionaries for instance (Riloff, 1993). Some are based on predefined knowledge like WordNet (Bagga 1997) to automate the production of pertinent patterns according to a relation, but such knowledge input is not directly usable (according to the domain, a word may have various meanings, so users may first have to select pertinent words and meanings according to their needs).

We have studied three learning methods adapted to extract information from texts, based on linguistic approaches:

- Rapier (Califf, Mooney 2003) learns rules or extraction patterns from an annotated learning corpus. It is based on syntactical analysis, on an algorithm applying compression rules, and on rules weighting. Rules are formalized in a way not adapted for non linguists (they use three filler patterns containing semantic and part-of-speech tags).
- ExDisco (Yangarber 2000) identifies a set of relevant documents and a set of event patterns from un-annotated texts, starting from a small set of weighted "seed patterns". These seed patterns are expanded to identify most of the various forms of the language expressing each relation in the documents. A limitation of this method is the use of a large training corpus, necessary to efficiently weight documents and patterns.
- Prométhée (Morin 1999) incrementally learns a set of lexico-syntactical patterns expressing a semantic relation from a set of terms linked by this relation. It is based on the Hearst algorithm which consists of having a first set of couples verifying an interesting relation, and constructing the corresponding set of patterns expressing this relation. These patterns permit the identification of new couples, that bring back new patterns… Patterns are expressed with regular expression notations: "CN_1 is? a NP_2 company" (where CN represents company name and NP a noun phrase), which are not easy to manipulate for non-specialists.

2.3 Our approach for the strategic intelligence

We have chosen to exploit a learning method, which is a good compromise between the costly descriptive methods and the inaccurate statistical methods. Our aim is to adapt a learning method to the specific constraints of the intelligence domain: we can't learn from large corpus (they don't exist on a new event) neither from annotated corpus (annotation by linguists of confidential documents is not conceivable). For the same reasons, we can't use adapted ontology developed by a linguist or a terminologist. And finally, we can't provide a method requiring linguistic knowledge to users specialized in intelligence and not language.

2.3.1 Learning algorithm

Our algorithm is based on the Prométhée algorithm but adapted to the intelligence domain as previously explained. It is applied on a pre-tagged text with entities. 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 and the automatic transformation of these extracts into patterns with Intex.

5. Use of the patterns: recovery of new couples. Back to the step 2.

In order to simplify the reading of the corpus by the user and the extraction of the couples, the corpus was 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.

2.3.2 Example

To illustrate our approach, here is an example from a French corpus. From the following sentences: "… le président Chirac avait téléphoné mercredi au président ivoirien Laurent Gbagbo … Laurent Gbagbo a reçu hier en sa résidence de Cocody, l'ancien président Henri Konan Bédié …"[3] users will obtain the following relations: "Jacques Chirac – A CONTACTÉ – Laurent Gbagbo" and "Laurent Gbagbo – A RENCONTRÉ – Henri Konan Bédié"[4]. These relations will then be merged in their knowledge management system. To do so, the first pattern "Jacques Chirac avait téléphoné ** à Laurent Gbagbo" is created by the users from the reduced lemmatised corpus, with this additional information: "Jacques Chirac" is the agent, "A CONTACTÉ" is the relation expressed by the pattern and "Laurent Gbagbo" is the patient. Applied on another corpus containing "Jacques Chirac avait téléphoné le 17 juillet au président Kostunica …", this algorithm will automatically produce the new relation "Jacques Chirac – A CONTACTÉ – Vojislav Kostunica".

3 SEM+

We present here Sem+, the tool containing our learning algorithm, and its evaluation.

3.1 First implementation

The first version of the Sem+ system was developed by Julia Frigière in 2004 in Java (Frigière, 2004). It is based on Intex, a linguistic development environment (Silberstein), and is developed in Java and Perl.

Before the use of Sem+, the users had to construct their own domain dictionaries, compiling their lists of words to each category (one text file for each entity category). Lemma can be used to unify outputs.

Here are the steps for using of Sem+ (see the figure below).

1. Selection of a corpus, and automatic tagging of the named entities described in the dictionaries (for example, companies.dic contains Transiciel and Gencom).

2. Reduction of the corpus according to a couple of entity categories (Company in our example).

3. Learning step (the two sub-steps below could be applied constantly):
    a. Capture of the entity couples to initiate the learning approach / recovery of new couples obtained from the captured patterns.

[3] "the president Chirac had phoned on Wednesday to the president of the Ivory Coast Laurent Gbagbo … Laurent Gbagbo has received yesterday, in his residence of Cocody, the previous president Henri Konan Bédié".

[4] "Jacques Chirac – HAS CONTACTED – Laurent Gbagbo" and "Laurent Gbagbo – HAS MET – Henri Konan Bédié"

b. 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: "Transiciel rachète Gencom"[5]. The user validates the agent, the patient, and the category of the relation expressed in the extract. For each extract, a transducer is automatically created with Intex to obtain for example ACHAT(Company1, Company2)[6] from "Company1 rachète Company2".

4. Use of the pattern set to extract relations on new corpus: "eBay rachète iBazar" => ACHAT(eBay, iBazar).

5. Production of a file with the Ideliance format, in order to export these relations into the knowledge management system.

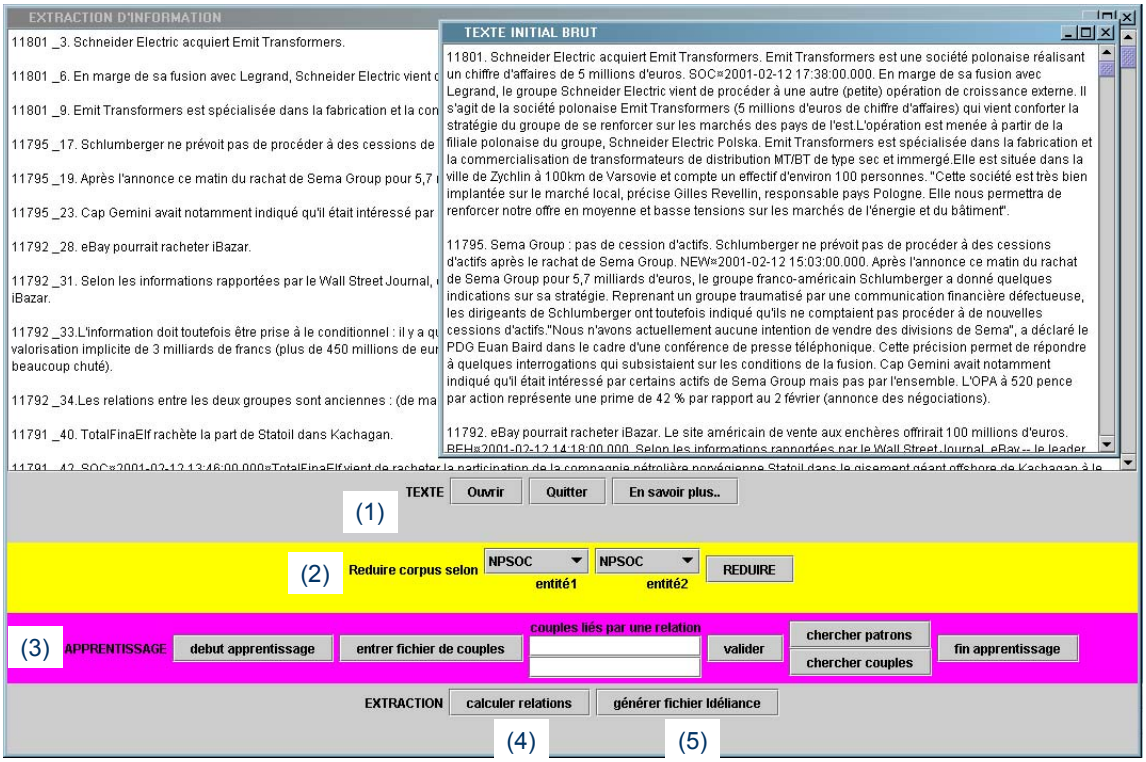

Fig. 1: Example of the Sem+ interface.

Sem+ and the underlying algorithm are adapted to various languages (English, French, …), as the patterns are developed from the corpus. They are also efficient for various users, because they can define their own entity categories, their own relations and select each pattern expressing them without any linguistic knowledge.

3.2 Evaluation

We have done two evaluations: the first was carried out on the initial corpus, which dealt with sales and purchasing relations between companies, and the second was performed on a more realistic corpus dealing with the Ivory Coast events, with seventeen relations between four entity categories.

[5] "Transiciel buys out Gencode."

[6] PURCHASE(Company1, Company2)

3.2.1 First evaluation on the financial corpus

The first evaluation (Frigière, 2004) was done on a financial corpus, to test the patterns coverage and the efficiency of the learning cyclical algorithm. There was only one entity category: Company, and two relations between companies: Sale, Purchasing.

On a sub-corpus containing 201 documents, Sem+ was used to capture 45 patterns. These patterns permitted the identification of 59 relations between companies (out of the 119 of this corpus). The recall was 41% and the precision was 98%. Several 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). Secondly, in the implemented version of Sem+, the patterns began with an entity and ended with the other entity, so patterns such as: “l’achat de Arisem par Thales” could not be exploited.

To evaluate the efficiency of the learning algorithm, 20 couples of companies verifying a relation of sale or purchasing were identified from the sub-corpus. They permitted the capture by the user of 28 patterns. These patterns were used to extract two other couples, supplying 4 new patterns. This result was motivating, as it illustrates the efficiency of our method to easily learn new patterns: 20 couples permitted the capture of 32 patterns.

3.2.1 Second evaluation on Ivory Coast corpus

The aim of this evaluation was to validate the usability of such an approach by a user without specific linguistic knowledge, and to test the method on another domain with a lot of relations and entity categories. This evaluation was done at the Land Headquarter (S.T.A.T. unit), by the Officer Cadet Cytermann (Cytermann, 2005). He used Sem+ on a corpus composed of journalistic articles related to the Ivory Coast crisis. For this evaluation, no specific criterion were used to quantify the results.

After some technical adjustment and a first test, it was concluded that Sem+ was easy to use, and that it was efficient in saving time when coupled with the Ideliance system. It was also remarked that Sem+ was efficient for manipulating a lot of entity categories and relations: for the evaluation, four entity categories were used: Person, Organization, Location and Mean, and seventeen relations: Meeting(Person, Person), Appointment(Person, Person), Support(Organization, Organization), Moving(Person, Location), etc.

On this corpus, the learning algorithm was not efficient, maybe because of the small size of the acquisition corpus (120 sentences). Also, there were a lot of relations to identify, so 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.

3.2.2 Entity tagging

The tagging step requires the use of dictionaries associated to each entity category. For example, on the financial corpus, we built a company name dictionary, containing lemmas (ex: British Telecommunications is associated to the lemma British Telecom). We also use a dictionary containing the potential initial sections of the company names, defined by T. Poibeau (Poibeau, 2003): “le groupe Thales”, “la société Thales”[7] are then reduced into “Thales”.

For the evaluation on the Ivory Coast corpus, the user also defined a dictionary for each entity category. But, as there was no dictionary to manage the potential initial sections, they were included in the

[7] Thales group, Thales company

dictionary: “le président Jacques Chirac”[8] was an input of the person name dictionary. To improve this, we have to efficiently manage these words that are parts of the person description.

Another difficulty was encountered: France was an element of the location dictionary, but in a lot of sentences its signification was more “the organisation of the country” than “the country”, as in: “La France accuse Gbagbo”[9]. The use of a location word to express a location or the associated organisation must be managed specifically in order to improve the results of our system.

4 CONCLUSION

We have worked on the automatic extraction of relations from texts. To do so, we have defined a specific algorithm, and implemented it with the Sem+ tool. Sem+ is easy to use for users without linguistic knowledge. It eases the reading of a corpus by reducing it according to entity categories pertinent to a specific relation. It also capitalizes the patterns expressing the relations, to reuse them for the automatic extraction of new relations.

As detailed previously, some evaluations were done to identify the interest of the approach to extract relations in an intelligence context, as well as to identify the improvements that must be provided. First, we want to simplify the use of the Sem+ tool, and thanks to the remarks made by the Officer Cadet Cytermann, we have listed some elements to improve: the management of the entity dictionaries must be done from the Sem+ interface; pre-filling of some information may improve the user task, etc. Secondly, we want to improve the learning algorithm efficiency, by improving the quality of the prior entity tagging step, by enlarging the patterns with the use of lemmas, by permitting the use of patterns beginning with a noun (“l’achat de Arisem par Thales”[10]), and by studying the characteristics of the acquisition corpus. As we stated, users don’t have large corpus on their intelligence subjects, but they can use a corpus on a similar topic, which may contain the pertinent relations. The use of lemmas will necessitate the use of a generic dictionary of the language: with the Intex system, we have such dictionaries for French and English. For the other languages, the use of lemmas will be constrained by the possession of the adequate dictionary.

As we said previously, Sem+ automatically provides results that are usable by the Ideliance system. We are also working with another team in THALES Research & Technology to provide results for a specific fusion module (Laudy, 2005).

We recently experimented with Sem+ in an audio filtering platform: relation patterns were used to extract relations from audio transcriptions of English audio data. This experiment showed the usability of Sem+ on English data, and a new efficient way to exploit the patterns defined from text to extract information from audio data.

Acknowledgement
We would like to thank the Lieutenant Colonel de Nicola and the Officer Cadet Cytermann from the S.T.A.T.

[8] The President Jacques Chirac

[9] France accuses Gbagbo.

[10] The purchase of Arisem by Thales

BIBLIOGRAPHY

Aussenac-Gilles N, Biébow B, Sulzman S (2000), Corpus analysis for conceptual modelling, in Proceeding of the 12th International Conference on Knowledge Engineering and Knowledge Management (EKAW'2000), pp13-20.

Bagga A., Chai J. Y. (1997) A Trainable Message Understanding System, Proceedings of ACL Workshop on Computational Natural Language Learning (CoNLL'97), Madrid, p.1-8.

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.

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

Desclés J-P. (1997), Systèmes d'Exploration Contextuelle, dans Co-texte et calcul du sens, Claude Guimier (éd).

Frigière J. (2004), Information extraction by learning method, Thales report.

Harris Z. S. (1968), Mathematical structures of language, Wiley, New York.

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

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.

Poibeau T. (2003), Extraction automatique d'information, du texte brut au web sémantique, Hermes.

Riloff E. (1993), Automatically Constructing a Dictionary for Information Extraction Tasks, in Proceedings of the Eleventh National Conference on Artificial Intelligence, 811-816. AAAI Press / The MIT Press.

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.

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

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.