First experiments of using semantic knowledge learned by ASIUM for information extraction task using INTEX

David Faure[1] and **Thierry Poibeau**[2]

Abstract. Our aim in this article is to show how semantic knowledge learned for a specific domain can help the creating of a powerful information extraction system. We describe a first experiment of coupling an information extraction system based and the machine learning system ASIUM. We will show how semantic knowledge learned by ASIUM helps the user to write an information extraction system more efficiently, in reducing the time spent on the development of resources. Our approach will be compared to the European ECRAN project, that aims at the same result, regarding development time and performances.

1 Introduction

Information Extraction (IE) is a technology dedicated to the extraction of structured information from texts. This technique is used to highlight relevant sequences in the original text or to fill pre-defined templates [1]. Below is the example of a story concerning a terrorist attack in Turkey together with the corresponding entry in the database filled by the IE system.

```
940815LM347810 Le Monde - 15 août 1994, page 6
TURQUIE: neuf blessés dans un attentat à la bombe.
Neuf personnes, dont trois touristes étrangers, ont été blessées par l'explosion
d'une bombe vendredi 12 août dans une gare routière de la partie européenne
d'Istanbul. (...)- (AFP.)

940815LM347810 Le Monde - August 15, 1994, page 6
TURKEY: nine persons were injured during a bomb attack.
Nine persons, three of them being foreign tourists, were injured by a bomb
explosion on Friday August 12, at a bus station in the European part of Istanbul.
```

Date of the story	15 août 1994
Loc.	TURQUIE Istanbul
Date	Vendredi 12 août
Nb dead person	
Nb person injured	neuf (nine)
Weapon	bombe (bomb)

Even if IE seems to be now a relatively mature technology, it suffers from a number of yet unsolved problems that limit its dissemination through industrial applications. Among these limitations, we can consider the fact that systems are not really portable from one domain to another. Even if the system is using some generic components, most of its knowledge resources are domain-dependent. Moving from one domain to another means re-developing some resources, which is a boring and time-consuming task[3].

This fact was observed by one of the authors during the elaboration of a previous prototype in the framework of the European ECRAN project [3] with the same aim. The system necessitated resources manually defined from the reading of a huge amount of texts.

In order to decrease the time spent on the elaboration of resources for the IE system, we suggest to use ASIUM that allows to learn semantic knowledge from texts. This knowledge is then used for the elaboration of the IE system. We also aim at reaching a better coverage thanks to the generalization process implemented in ASIUM.

We will firstly present the ASIUM system which allows to learn semantic knowledge for the elaboration of an IE system similar to that of the ECRAN project. We will show to what extent it is possible to speed up the elaboration of resources without any decrease of the quality of the system. We will finish with some comments on this experiment and we will show how domain-specific knowledge acquired by ASIUM such as the subcategorization frame of the verbs could be used to extract more precise information from texts.

2 Semantic Knowledge Acquisition

Semantic knowledge acquisition from texts remains a hard task even for limited domains. This knowledge is crucial in order to improve natural language applications like information extraction. Approaches mixing machine learning (ML) and natural language processing (NLP) obtain good results in a short development time (we can cite, among others M. E. Califf [4], R. Basili [5], S. Buchholz [6], D. Hindle [7], R. J. Mooney [8] et E. Riloff [9], [2], [10]).

We present here ASIUM which learns cooperatively semantic knowledge from texts syntactically parsed without previous manual processing. This knowledge consists in subcategorization frames of verbs and an ontology of concepts for a specific domain following the "*domain dependence*" defined by G. Grefenstette[4] [11].

ASIUM is based on an unsupervised conceptual clustering method and provides an ergonomic user-interface[5] to help knowledge acquisition process.

In this part, we will show how ASIUM is able to learn good quality knowledge in a reasonable time from parsed text, even if the syntactic parsing of texts is noisy.

2.1 Our approach

Our aim is to learn subcategorization frames of verbs and an ontology for a specific domain, from texts. Actually, existing knowledge bases like EUROWORDNET or WORDNET are frequently over-general for applications in specific domains. These ontologies, although very complete, are not suitable for processing texts in technical languages. On one hand they are not purpose directed ontologies, they may store up to seven meanings and syntactic roles for a word, thus increasing the risk of semantic ambiguity. In a specific domain, the vocabulary as well as its possible usage is reduced, which makes ontologies such

[1] L.R.I. UMR 86-23 du CNRS Université Paris Sud F-91405 Orsay Cedex David.Faure@lri.fr

[2] Thomson-CSF Laboratoire Central de Recherches Domaine de Corbeville F-91404 Orsay Thierry.Poibeau@lcr.thomson-csf.com

[3] for example Riloff [2] mentions a 1500 hours development.

[4] "*A semantic structure developed for one domain would not be applicable to another*".

[5] http://www.lri.fr/Francais/Recherche/ia/sujets/asium.html

as WORDNET overly general. On the other hand, WORDNET may lack some specific terminology of the application domain.

Contrary to any approach of increasing or specializing general ontologies for a specific domain like R. Basili [5], we learn an ontology and verbs frames from the corpus reducing the risk of inconsistency.

Our previous attempts to automatically revise subcategorization frames and a subset of an ontology acquired by a domain expert have failed. Revision of the acquired knowledge with respect to the training texts required deep restructuring of the knowledge that incremental and even cooperative ML revision methods were not able to handle. The main reason was that the expert built the ontology and the subcategorization frames with too many *a priori* that were not reflected in the texts. This experiment illustrates one of the limitation of manual acquisition by domain experts without linguists.

2.2 Learned knowledge

ASIUM learns subcategorization frames like `<to drop> <object: Explosive> <in: Public_Place>` for the verb `to drop`. Both couples `object: Explosive` et `in: Public_Place` are *subcategories*, `object` is a *syntactic role* and `in` is a *preposition* but `Explosive` and `Public_Place` are concepts used as *restrictions of selection*. More usually, ASIUM learns verb frames like: `<verb> <prep.|syntactic role: concept*>*`

These frames are more general than the ones defined in the LFG [6] formalism because the subcategories are verb arguments (subject, direct object or indirect object) and adjuncts. In our framework, restrictions of selection can be filled by an exhaustive list of nouns (in canonical form) or by one or more concepts defined in an ontology. The ontology represents generality relations between concepts in the form of a directed acyclic graph (DAG). For example, the ontology could define `car`, `train` and `motorcycle` as `motorized vehicle`, and `motorized vehicle` as both `vehicle` and `pollutant`. Our method learns such an ontology and subcategorization frames in an unsupervised manner[7] from texts in natural language. The concepts formed have to be labeled by an expert.

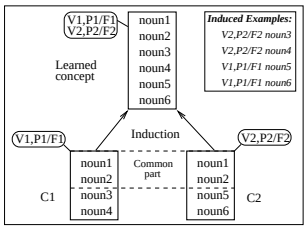
2.3 Knowledge acquisition method

The first step of the acquisition process is to automatically extract syntactic frames from texts. We use the syntactic parser SYLEX developed by P. Constant [12]. In case of syntactic ambiguities, SYLEX gives all the differents interpretations and ASIUM uses all theses interpretations. Experiments have shown that the ML method works well with theses ambiguities and acquisition of semantic knowledge is not affected. This method avoids a very time-consuming manual disambiguation step. These frames are the same like subcategorization frames but with concepts replaced by nouns. `<verb> <prep. | role: head noun>*`

ASIUM only uses head nouns of complements and links with verbs. Adjectives and empty nouns are not used. Our experiments have shown that these informations were enough to learn semantic knowledge even from a noisy syntactic parsing.

The learning method relies on the observation of syntactic regularities in the context of words [13]. We assume here that head nouns occuring with the same couple `verb+preposition/syntactic role` represent a so-called *basic class* and have a semantic similarity in the same line as Grefenstette[11], Peat[14] or others, but our method is based on a *double regularity model*: ASIUM gathers nouns together as representing a concept only if they share at least *two different* `(verb+preposition/syntactic role)` contexts as in Grishman[15]. Experiments show that it forms more reliable concepts, thus requiring less involvement from the user. Our similarity measure computes the overlap between two lists of nouns [8] (Details in [16]). As usual in conceptual clustering, the validity of learned concepts relies on the quality of the similarity measure between clusters that increases with the size of their intersection.

Basic classes are then successively aggregated by a bottom-up breadth-first conceptual clustering method to form the concepts of the ontology level by level with an expert validation and/or labelling at each level. Thus a given cluster cannot be used in a new construction before it has been validated. For complexity reasons, the number of clusters to be aggregated is restricted to two, but this does not affect the relevance of the learned concept [16]. Verb subcategorization frames are learned in parallel so that each new concept fills the corresponding restriction of selection then resulting in the generalization of the initial syntactic frames which allows to cover examples that *did not occur* as such in texts. Thus, the clustering process does not only identify the lists of nouns occuring after the same `verb+preposition/function` but also augments this list by *induction*.



Aggregation of two basic classes (C1 and C2) found after two different couples `verb+prep./function` (V1,P1/F1 and V2,P2/F2) will create a new concept allowed after V1,P1/F1 and V2,P2/F2. Thus, nouns which only appear in basic class C1 (resp. C2) will now be allowed with the couple V2,P2/F2 (resp. V1,P1/F1). This results in a generalization of knowledge found in the corpus as presented in the figure.

For example, starting with these syntactic frames,

- `<to travel>`
  `<subject:[father,neighbour,friend]>`
  `<by: [car,train]>`
- `<to drive>`
  `<subject:[friend,colleague]>`
  `<object:[car,motorcycle]>`

ASIUM will learn two concepts

- `Human`: `father`; `neighbor`; `friend`; `colleague`.
- `Motorized Vehicle`: `car`; `train`; `motorcycle`.

and two subcategorization frames:

- `<to travel>`
  `<subject: Human>`
  `<by:Motorized Vehicle>`
- `<to drive>`
  `<subject: Human>`
  `<object: Motorized Vehicle>`

Experts have to control the link between the new concept and the verb because the only threshold, fixed by the expert, can not measure the over-generalization risk. This validation process is relatively quick due to the ergonomic user-interface. ASIUM provides to the expert the list of newly covered examples in order to estimate the generality of the proposed concept. Moreover the expert can use functionalities provided by ASIUM in order to divide the learned concept into sub-concepts in case of a proposed concept overly general for the target task.

[6] Lexical Functional Grammar.

[7] Asium is called unsupervised because no concepts examples are provided as input.

[8] $Sim(C_1, C_2) = 1$ for lists with the same nouns and $Sim(C_1, C_2) = 0$ for lists without any common nouns.

2.4 Related work in semantic knowledge acquisition

As for D. Hindle [7] or F. Peireira [17], our method gather nouns regarding syntactic regularities of arguments and adjuncts of the verbs. We suppose that in specialized texts, verbs are also characterized by theirs adjuncts. G. Grefenstette [11] proposes to learn something close to our "*basic classes*". Our "*double similarity model*" learns a concept by gathering two basic classes only if they have a good similarity. This model limits the number of non relevant produced concepts. M. R. Brent [18] learns only five subcategorization frames from untagged texts with an automatic method. S. Buchholz [6] learns subcategorization frames very close to ours but with a supervised method which is very time-consuming for the expert. In the same way, WOLFIE (A. C. Thompson [19]) with CHILL (J. M. Zelle [20]) learns "*case-roles*" and a thesaurus from texts syntactically parsed by CHILL but fully semantically annotated by hand. The case roles differs from our subcategorization frames because our prepositions or grammatical functions are replaced by semantic roles like *agent* or *patient*. Contrary to the ontology learned by ASIUM, selectional restrictions learned by WOLFIE are attribute-value lists. An unsupervised learning approach like ASIUM delays concepts labelling after the learning process and so considerably reduces the time needed by the expert. After ASIUM learning, the semantic roles can be labelled by assuming a couple `verb+prep./function` represents a specific semantic role. E. Riloff in [10] learns five concepts from texts. She uses lists of nouns representing general concepts (seeds) and uses coocurrence method to augment these lists to concepts. These augmented lists are checked by the expert who only retains nouns representing the concept. We can assume basic classes of ASIUM are seeds that will be increased by our induction process. The main advantage is that the number of concepts is not limited to five and we learn in parallel subcategorization frames of verbs without more time-consuming validation needed.

3 The Information Extraction system

The Information Extraction system is based on the INTEX tool-box, developed by the LADL laboratory[9]. INTEX allows a rapid and interactive development of automata and transducers to analyze texts. A linguistic automaton recognizes expressions in texts, whereas a transducer associate specific tags with words in the texts (for example, assign a syntactic category to a word). Transducers are efficient, expressive and sufficient for a local analysis of texts. We chose this approach because it allows the rapid development of an IE system for a given domain with a strictly local analysis limited to the sentence area. Our aim is to develop a highly portable system even if this means using more precise analysis strategies afterwards.

3.1 Linguistic resources modeling

To elaborate linguistic resources, we first used the semantic classes defined by the ASIUM system. Before the experiment, the corpus was separated in two different parts : the training set and the test set. The linguistic resources are constantly tested on the training set during the development. This development approach allows to evaluate the performances and to detect possible errors in the grammar (a grammar with too much or not enough constraints which would bring silence or noise during the analysis). The expressions modeled via transducers are for most part syntactic structures (the set of expressions equivalent to the notion of "`bombing`") integrating some of the semantic classes furnished by the ASIUM system [10].

The homogeneous semantic lists learned by the ASIUM system are introduced in the INTEX vocabulary. At this level, a manual work is necessary to exploit the semantic classes from ASIUM. These classes are refined (merging of scattered classes, deletion of irrelevant elements, addition of new elements, etc.). About ten hours have been dedicated, after the acquisition process, to the refinement of the data furnished by ASIUM. This knowledge is then considered as a resource for INTEX and is exploited either as dictionaries or as transducers, in function of the nature of the information. If it is a general information that is not domain specific, we prefer to use a dictionary which can be reused, otherwise, we use a transducer.

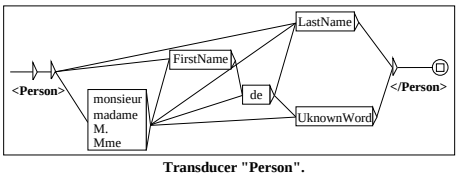
A dictionary is a list of words or phrases, each one being accompanied of a tag and a list of features[11]. The first names dictionary or the locations dictionary are generic reusable resources. Below is a sample of the location names dictionary[12]:

Abidjan,*N+Loc+City*;
Afghanistan,*N+Loc+Country*;
Allemagne,*N+Loc+Country*;
Allemagne de l'Ouest,*N+Loc+Country*;
Allemagne de l'Est,*N+Loc+Country*; . . .

These items structured in a list are convenient for the dictionary format and the semantic lists elaborated from ASIUM complete in an accurate manner the coverage of the initial dictionaries from INTEX.

The transducer format is essentially used for more complex or more variable data where linguistic phenomena such as insertion or optionality may interfere.

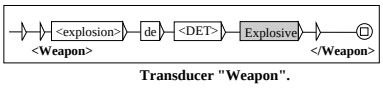
Here, the figure presents an example of a transducer allowing to recognize person names such as *Monsieur Jean Dupont*. the transducer recognizes a sequence composed of a trigger word (*Monsieur*), a first name (*Jean*) and a proper name (*Dupont*). But we must keep in mind that most of these elements can be optional (*Monsieur Dupont* or *Jean Dupont* are correct sequences) and that *Dupont* can be a word that is not listed in any dictionary (it will then be considered as an unknown word).



Transducer "Person".

At this level, one can find two types of transducers: some are generic - as the "Person" one, and some others are domain-specific and can filled with the semantic knowledge acquired by ASIUM.

The next figure is the illustration of a transducer recognizing `explosion de Det N` (explosion of Det N), where the nominal phrase `Det N` recognizes nominal phases elaborated from the semantic class `bombing` where the following words appear: `bombe (bomb)`, `obus (shell)`, `grenade`, etc.



Transducer "Weapon".

The elaboration of such transducers requires some linguistic expertise to obtain *in fine* a system recognizing the relevant sequences without too much noise. The architecture of the system is using cascading transducers, it is then important that each level has a good quality in order to allow the following analysis level to operate on a solid background.

[9] Laboratoire d'Automatique Documentaire et Linguistique de l'Université de Paris 7.

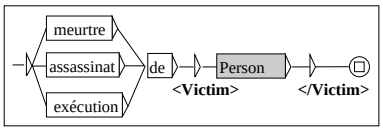
[10] for example the list of weapons which could be used in a bomb attack.

[11] For example, in this case: *Loc,City,Country* . . .

[12] Each line begins with a term, followed by some indication about its syntactic category (`N` for `Noun`) and semantic features (`Loc` to indicate a location, `Country` to indicate a country, etc.).

This kind of architecture amplifies indeed systematically the noise generated by the previous level.

For example, the results of the transducer presented on right figure[13] will be better if those of the transducer "Person" (in grey) are already good.



Transducer "Victim".

The different defined transducers are then minimized and determined [14]. The overall set of transducers is composed of 1000 nodes and about 5000 arrows in our experiment.

3.2 Related work in Information Extraction

IE is a now widely spread research domain. The American Message Understanding Conferences (MUC) provided a formidable framework for the development of research in this area ([21], [22]). The conferences are held about every two years and generally bring together about fifteen teams working on IE systems. The elaboration of the linguistic resources is for most part a manual work even if some attempts were done to have some more portable systems.

At least two French-speaking projects have been developed which are somewhat comparable with MUC systems. these two systems are the European project ECRAN and the EXIBUM project from the University of Montreal (Canada). ECRAN developed a generic and multilingual system tested on different corpora (movie reviews, stories from the economic area, etc.) [23]. EXIBUM is a bilingual system (French and English) that aims at processing agency news about terrorist events in Algeria [24].

Several other Information Extraction systems were developed for specific kind of information (dates, location names, etc.). For example, D. Maurel [25] developed a system highlighting dates by means of automata and acceptability tables. More recently, C. Belleil [26] presented a system highlighting French toponyms and J. Sénellart [27] a system recognizing Minister names from the French newspaper *Le Monde*. These Approaches generally require exhaustive descriptions of the concerned domain.

Recent American work in the area proposed an approach mixing corpus exploration and knowledge acquisition to feed IE systems. A first well-known experiment is the AutoSlog from E. Riloff [2] allowing to find in texts relevant syntactic structures from keywords given to the system by the end-user. In the framework of ECRAN, a similar attempt was done to try to generalize relevant syntactic structures from a training corpus and a general dictionary [28]. The experiment we present is different considering that the learning system is not supervised and furnishes the IE system designer a wide amount of knowledge extracted from the texts.

4 Experiment

In our experiment, we have used a corpus of texts form the French journal "*Le Monde*". Texts indexed by the noun "*terrorist event*" have been extracted and manually filtered in order to be sure that they really contain a terrorist event description[15]. This corpus is of the same kind as the one used for experiments in the ECRAN project, so that we will be able to compare our results.

The time spent on the definition of the linguistic resources with INTEX is estimated to about 15 hours. This duration has to be compared with the two weeks[16] needed for the manual resources development of the ECRAN project.

Hundred texts have been used as "training corpus" and fifteen different texts have been used as "test corpus". Texts are first parsed with our system, and then some heuristics allow to fill the extraction template:

- Due to the structure of articles of *Le Monde*, the first date is always the date of the article;
- we assume that the second date is the one of the terrorist event;
- the two first occurrences of locations found are stored and usually quite well identify the location of the terrorist event;
- the first occurrence of a number of victims or injured persons is stored. If a text speaks of more than one terrorist event, we assume that only the first one is relevant. We have chosen short texts to prevent us from this problem inherent to long texts;
- only the first weapon linked with the terrorism event is stored.

These heuristics are very succinct and we will have to specialize them to perform information extraction on longer or less-specialized texts. We have used these simple heuristics to evaluate our system and compare it with the ECRAN one. With these heuristics, we obtain good results on our corpus, and most of the extraction systems evaluated in the American MUC conferences used this kind of heuristics in order to solve any parsing problems.

Our results have been evaluated by two human experts who did not follow our experiment. Our performance indicators were defined as:

- OK (O) if extracted information is correct;
- FALSE (F) if extracted information is incorrect or not filled;
- NONE (N) if there were no extracted information and no information has to be extracted.
- FALSE for all the other cases.

Using these indicators, we can compute two differents values:

- PRECISION1 (P1), ratio between OK and FALSE answers, without taking into account the NONE answers.
- PRECISION2 (P2), same as P1 but with the NONE answers.

The next table summarizes results for the different elements of the template.

	O	F	N		P1	P2
Date of the story	50	0	0		1,00	1,00
Location	45	5	0		0,90	0,90
Date	49	1	0		0,98	0,98
Nb dead persons	20	5	25		0,80	0,90
Nb persons injured	26	9	15		0,74	0,82
Weapon	35	11	4		0,76	0,78
Average	37,5	5,2	7,3		0,86	0,89

We obtain a good quality for the extracted information in most of the elements.

- The date of the story is fully correct because we can use the structure of the article to extract it;
- The errors for the location slot are due to two "contradictory" locations found by the system. A more complete linguistic analysis or a database providing lists of cities in different countries would reduce this kind of errors;
- The errors in the number of dead or injured persons slot are frequently due to silence. Our system, for example, fails against too complex syntactic forms like "*Deux médecins italiens travaillant*

[13] meurtre=murder, assassinat=assassination, éxecution=execution.

[14] These two operations allow to optimize the analysis time.

[15] The full corpus also contains others texts describing proceedings or terrorist menaces.

[16] about 80 hours.

pour médecins sans frontières (MSF-Belgique) ont été blessés. (*Two Italian doctors working for médecins sans frontières (MSF-Belgique) have been injured.*)", where the passive subject have not been correctly parsed by the system;

• The silence for the weapon slot is frequently due to incompleteness of semantic dictionaries.

5 Discussion

In this section, we will comment some of the results of this experiment. Results obtained prove the interest of coupling a semantic knowledge acquisition tool with the IE system. But those results are not precise enough to decide about the quality of the semantic knowledge acquisition tool. We will examine here some indicators which allow to judge of the quality of the semantic knowledge learned and next we will present some comments on the information extraction.

5.1 Semantic Knowledge quality

Semantic knowledge acquisition tools like ASIUM are always very difficult to evaluate. Measuring the quality of an ontology or evaluating an ontology regarding another one is not easy and heavily depends on applications. So, we will only present here some indicators to have an idea on the quality of the acquired knowledge.

Concept quality depends of two different elements. The first one is the distance which computes similarity between classes in order to create relevant concepts and perform relevant inductions. As usual in conceptual clustering, the distance is a parameter of the concept quality and of quantity of expert's work.

This first qualitative element is very hard to estimate. In our application, 16 of the 19 first classes proposed by ASIUM have been accepted by the expert. 447 inductions have been proposed by ASIUM and 73 % of these inductions have been judged relevant by the expert.

The second element which affects the concepts quality is the level of generality for a concept. When ASIUM proposes a new concept, the expert has to decide from the generality of the concept whether it should be split or not. This work is easy for an expert because he has a very good knowledge of the final application.

For example, if ASIUM proposes the "`Organization`" concept, the expert has to decide if it is relevant for the task to identify subconcepts like "`Military org.`" and "`Politic org.`".

The generality level in the application highly depends on the subtlety of the template to be filled by the information extraction system. Our previous experiments on this domain and on the cooking recipes domain have shown that this work is simple and that expert choices really depend on the task. (More explanations on the suitability of concepts for the main task and unsuitability of these concept for another task are given in [29].)

5.2 Comments on the extraction process

The results we obtained during this experiment can be satisfactorily compared with those that we obtained on the same corpus with the ECRAN system, that performed 0.89 precision. Moreover, the results we performed with the new system were obtained after a reduced development phase: about 40 hours for the learning phase with ASIUM and about 15 hours to format the knowledge base as INTEX resources. The following comments can be done on this experiment:

• Having a good knowledge of the corpus is indubitably an advantage for the system designer. The fact that one of the author had previously done the same task for ECRAN speeded up the development process, given that the search of relevant syntactic structures was facilitated;

• The results of the ASIUM system allow to speed up the definition of the paradigmatic classes filling states in the INTEX transducers, even if certain classes need to be manually completed. For example, the ASIUM semantic classes allowed to rapidly complete the graph representing the set of weapons or persons who were implicated in terrorist events. ASIUM provided a class in which terms such as, for example: "`bomb`", "`grenade`", "`explosive`" or "`car`" could appear, considering that a *booby-trapped car* is a kind of weapon, etc;

• The description language provided by INTEX is richer than the one of ECRAN. The time spent to model the linguistic INTEX transducers was longer than the one spent for ECRAN since the constraints and the empty transitions in automata and transducers have to be manually designed so that the noise is kept at a low level [17].

Such an evaluation, in which we deliberately limited the time spent on the development of linguistic resources, shows the importance of having accurate resources adapted to the task. Moreover, the inescapable incompleteness of the developed resources facing new texts shows that this kind of systems have to integrate dynamic acquisition processes to assist the incremental enrichment of resources, as time goes by.

The experiment was intended to show the time needed for the development of a sufficient set of resources, in order to obtain results equivalent to those of the ECRAN project. That is the reason why we emphasize on an evaluation of the amount of time spent on the task rather than on the improvement potential. That is also the reason why we focused on a limited template that only necessitates a surface analysis. This limitation could certainly be solved if we used more accurately the knowledge acquired by ASIUM. Thus, we plan to take into account a deeper linguistic analysis (anaphora resolution, partial information merging, etc.).

6 Future work

All the knowledge learned by ASIUM is not used in this experiment, especially subcategorization frames. We showed that a surface analysis is sufficient when templates to be filled are not more complex than those of ECRAN. The good quality obtained in a very short time proves this idea.

Nevertheless, in order to extract more specific informations from texts (like the name of the organization that performs the terrorist event, the politic membership of victims or attacker nationality), we think that the use of subcategorization frames could be very useful. Writing syntactic rules in order to perform relevant information extraction becomes very hard because of the multiplicity of the syntactic variations used in texts.

Our current work is to create a cooperative acquisition system to learn resources using the subcategorization frames learned by ASIUM. The expert will be able to express rules using complements of verbs independently of the syntax. Active and passive forms will be given the same representation by the system. For example, the two following sentences will be equivalent: *L'action terroriste est revendiquée par le Front populaire de libération de la Palestine (FPLP) (The terrorist event was claimed by the FPLP)* or *le Front populaire de libération de la Palestine (FPLP) revendique l'action terroriste (The FPLP claimed responsibility of the terrorist event.)* One example of rules for this kind of sentences can be:

[17] The effort to manage empty transitions in graphs took about 5 hours but allowed to obtain a more efficient grammar than the one obtained by the description of syntactic patterns by a set of regular expressions.

If verb is "*to claim*", and **object** belongs to the class "*Attack*" **Then** the **subject** is the attacker.

This kind of rule allows to differentiate people claiming terrorism events like in *Un groupe terroriste libanais revendique l'attentat antisémite de Buenos-Aires (A lebanese terrorist group claim anti-semite attack in Buenos-Aires)* from an organization claiming for a right like in *"les fondamentalistes musulmans revendiquent le droit de vote (Muslim fundamentalists are claiming voting rights)*.

Semantic rules allow to make fine differences to accurately fill fine-grain slots. The two next rules fill the field "`Missile`" or "`Attacker`" regarding the concept (`Explosive` or `Person`) learned by ASIUM and used as **subject** of the verb `to kill`.

If verb = "*to kill*" and **subject** = Person **Then** the **subject** is the attacker.
If verb = "*to kill*" and **subject** = Explosive **Then** The **subject** is the missile used.

We can see that, even if syntactic parsers generate errors and ambiguities, ASIUM can check texts using the ontology and the subcategorization frames previously learned. Then, the information extraction process will process only on consistent sentences with subcategorization frames. This allows to detect some parsing errors.

The system we are thinking of will process two different parsing steps. First, we will use syntax and concepts learned by ASIUM to pre-fill the frame. Second, we will use our "conceptual rules" to fill more specifically the frame.

7 Conclusion

We have described in this article an experiment in which we coupled an information extraction system using INTEX with the machine learning system ASIUM. The development time of the linguistic resources of the information extraction system has been reduced by using the semantic knowledge learned by ASIUM. The quality of the results remains the same as in the European ECRAN project.

The aim of this experiment was to validate our approach. We will now explore a better integration of the two systems and examine how to better use the semantic knowledge learned by ASIUM in order to increase the quality of our results.

Acknowledgment

The research from Thierry Poibeau is partially funded by a Cifre grant between the Laboratoire Central de Recherches of Thomson-CSF and the Laboratoire d'Informatique de l'Université de Paris-Nord.

The authors want acknowledge M. Rodde (Cristal-Gresec) and A. Balvet (Université Paris X) for their contribution during analysis of the results.

REFERENCES

[1] M. T. Pazienza, ed., *Information extraction (a multidisciplinary approach to an emerging information technology)*. Berlin: Springer Verlag (Lecture Notes in computer Science), 1997.

[2] E. Riloff, "Automatically generating extraction pattern from untagged texts," in *13th Conference on Artifcial Intelligence (AAAI'96)*, (Portland, Canada), 1996.

[3] T. Poibeau, "Mixing technologies for Intelligent Information Extraction," in *Proceedings of the workshop on Intelligent Information Integration, 16th International Joint Conference on Artificial Intelligence*, pp. 116–121, 1999.

[4] M. E. Califf, *Relational Learning Techniques for Natural Language Information Extraction*. PhD thesis, Department of Computer Sciences, University of Texas at Austin, February 1997.

[5] R. Basili and M. T. Pazienza, "Lexical Acquisition for Information Extraction," in *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology* (M. T. Pazienza, ed.), (Frascati, Italy), LNAI Tutorial, Springer, July 1997.

[6] S. Buchholz, "Distinguishing Complements from Adjuncts using Memory-Based Learning," in *Proceedings of the ESSLLI'98 workshop on Automated Acquisition of Syntax and Parsing* (B. Keller, ed.), pp. 41–48, 1998.

[7] D. Hindle, "Noun classification from predicate-argument structures," in *Proceedings of the 28st annual meeting of the Association for Computational Linguistics, ACL, Pittsburgh, PA*, pp. 1268–1275, 1990.

[8] R. J. Mooney, A. C. Thompson, and R. L. Tang, "Learning to Parse Natural Language Database Queries into Logical Form," *Proceedings of the ML-97 Workshop on Automata Induction, Grammatical Inference, and Language Acquisition*, 1996.

[9] E. Riloff, "Automatically Constructing a Dictionnary for Information Extraction Tasks," in *Proceedings of the Eleventh Nationnal Conference on Artificial Intelligence*, pp. 811–816, 1993.

[10] E. Riloff and J. Shepherd, "A Corpus-Based Approach for Building Semantic Lexicons," in *Proceedings of the Second Conference on Empirical Methods in Natural Language Processing (EMNLP-2)*, 1997.

[11] G. Grefenstette, "Sextant: exploring unexplored contexts for semantic extraction from syntactic analysis," in *Proceedings of the 30st annual meeting of the Association for Computational Linguistics, ACL*, (Newark, Deleware, USA), pp. 324–326, June 1992.

[12] P. Constant, "Reducing the complexity of encoding rule-based grammars," December 1996.

[13] Z. Harris, *Mathematical Structures of Language*. New York: Wiley, 1968.

[14] H. J. Peat and P. Willet, "The limitations of term co-occurrence data for query expansion in document retrieval systems," *Journal of the American Society for Information Science*, vol. 42, no. 5, pp. 378–383, 1991.

[15] R. Grishman and J. Sterling, "Generalizing Automatically Generated Selectionnal Patterns," in *Proceedings of COLLING'94 15th International Conference on Computational Linguistic*, (Kyoto, Japan), August 1994.

[16] D. Faure and C. Nédellec, "A Corpus-based Conceptual Clustering Method for Verb Frames and Ontology Acquisition," in *LREC workshop on Adapting lexical and corpus ressources to sublanguages and applications* (P. Velardi, ed.), (Granada, Spain), pp. 5–12, May 1998.

[17] F. Pereira, N. Tishby, and L. Lee, "Distributional Clustering of English Words," in *Proceedings of the 31st annual meeting of the Association for Computational Linguistics, ACL*, pp. 183–190, 1993.

[18] M. R. Brent, "Automatic acquisition of subcategorization frames from untagged text," in *Proceedings of the 29st annual meeting of the Association for Computational Linguistics, ACL*, pp. 209–214, 1991.

[19] C. A. Thompson, "Acquisition of a Lexicon from Semantic Representations of Sentences," in *33rd Annual Meeting of the Association of Computational Linguistics, Boston, MA July, (ACL-95).*, pp. 335–337, 1995.

[20] J. M. Zelle and R. J. Mooney, "Learning semantic grammars with constructive inductive logic programming," *Proceedings of the Eleventh National Conference on Artificial Intelligence*, pp. 817–822, 1993.

[21] "MUC-6," in *Proceedings of the sixth Message Understanding Conference (MUC 6)*, (San Francisco), Morgan Kaufmann, 1996.

[22] "MUC-7," in *Proceedings of the seventh Message Understanding Conference*, (San Francisco), Morgan Kaufmann, 1998.

[23] T. Poibeau, "Extraction d'information : adaptation lexicale et calcul dynamique du sens," in *Actes des rencontres internationales sur l'extraction, le filtrage et le résumé automatique (Rifra'98)*, (Sfax, Tunisia), pp. 141–153, November 1998.

[24] L. Kosseim and G. Lapalme, "EXIBUM : un système expérimental d'extraction bilingue," in *Actes des rencontres internationales sur l'extraction, le filtrage et le résumé automatique (Rifra'98)*, (Sfax, Tunisia), pp. 129–140, November 1998.

[25] D. Maurel, *Reconnaissance des séquences de mots par automate, adverbes de date du français*. PhD thesis, Université Paris 7, 1989.

[26] C. Belleil, *Reconnaissance, typage et traitement des coréférences des toponymes français et de leurs gentilés par dictionnaire électronique relationnel*. PhD thesis, Université de Nantes, 1997.

[27] J. Sénellart, "Locating noun phrases with finite state transducers," in *15th International Conference on Computational Linguistics (COLING'98)*, (Montréal), pp. 1212–1217, 1998.

[28] R. Basili, R. Catizone, M. T. Pazienza, M. Stevenson, P. Velardi, M. Vindigni and Y. Wilks, "An empirical approach to Lexical Tuning," in *Actes du Workshop on Adapting lexical and corpus resources to sublanguages and applications*, (Granada, Spain), May 1998.

[29] D. Faure, "Connaissances sémantiques acquises par Asium: exemples d'utilisations," in *Journée du Réseau de sciences cognitives d'Ile-de-France* (RISC, ed.), p. 12, October 1999.