

# Generating extraction patterns from a large semantic network and an untagged corpus

Thierry POIBEAU  
Thales and LIPN  
Domaine de Corbeville  
91404 Orsay, France  
Thierry.Poibeau@thalesgroup.com

Dominique DUTOIT  
Memodata and CRISCO  
17, rue Dumont d'Urville  
Caen, France  
memodata@wanadoo.fr

## Abstract

This paper presents a module dedicated to the elaboration of linguistic resources for a versatile Information Extraction system. In order to decrease the time spent on the elaboration of resources for the IE system and guide the end-user in a new domain, we suggest to use a machine learning system that helps defining new templates and associated resources. This knowledge is automatically derived from the text collection, in interaction with a large semantic network.

## 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 (Pazienza, 1997).

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 (for example Riloff

(1995) mentions a 1500 hours development). Several recent works propose to overcome these limitations by using annotated corpora as a reservoir of knowledge. However, annotated corpora are rarely present in companies, and to a certain extent solutions based on corpora seem to be inappropriate.

In this paper, we propose an approach based on a rich semantic network. We will firstly describe this network and a set of original measures we have implemented to calculate similarities between words. We will then present the acquisition process, in which the semantic network is projected on the corpus to derive extraction patterns. This mechanism can be seen as a dynamic lexical tuning of information contained in the semantic network. In the last section, we propose an evaluation and some perspectives.

## 2 Related work

The bases of IE as defined in the introduction are exposed in (Pazienza, 1997). IE is known to have established a now widely accepted linguistic architecture based on cascading automata and domain-specific knowledge (Appelt *et al*, 1993). However, several studies have outlined the problem of the definition of the resources, see E. Riloff (1995).

To address this problem of portability, a recent research effort focused on using machine learning throughout the IE process (Muslea, 1999). A first trend was to directly apply machine learning methods to replace IE components. For instance, statistical methods have been successfully applied to the named-

entity task. Among others, (Bikel *et al.*, 1997) learns names by using a variant of hidden Markov models.

Another research area trying to avoid the time-consuming task of elaborating IE resources is concerned with the generalization of extraction patterns from examples. (Muslea, 1999) gives an extensive description of the different approaches of that problem. Autoslog (Riloff, 1993) was one of the very first systems using a simple form of learning to build a dictionary of extraction patterns. Successors of AutoSlog like Crystal (Soderland *et al.*, 1995) mainly use decision trees and relational learning techniques to learn set of rules during their extraction step. More recently, the SrV system (Freitag, 1998) and the Pinocchio system (Ciravegna, 2001) use a combination of relational and basic statistical methods inspired from Naïve Bayes for IE tasks.

These approaches acquire knowledge from texts but they must be completed with a semantic expansion module. Several authors have presented experiments based on Wordnet (Bagga *et al.*, 1996).

Our approach is original given that it consists in an integrated system, using both a semantic network and a corpus to acquire knowledge and overcome the limitations of both knowledge sources. On the one hand, the fact that we use a semantic network allows us to obtain a broader coverage than if we only used a training corpus (contrary Ciravegna's system for example). On the other hand, the corpus ensures that the acquired resources are quite adapted to the task (contrary Bagga's system for example). The performance of the system will demonstrate this point (see below section 5).

### 3 The semantic net

The semantic network used in this experiment is a multilingual net providing information for five European languages. We quickly describe the network and then give some detail about its overall structure.

#### 3.1 Overall description

The semantic network we use is called *The Integral Dictionary*. This database is basically structured as a merging of three semantic

models available for five languages. The maximal coverage is given for the French language, with 185.000 word-meanings encoded in the database. English Language appears like the second language in term of coverage with 79.000 word-meanings. Three additional languages (Spanish, Italian and German) are present for about 39.500 senses.

These smallest dictionaries, with universal identifiers to ensure the translation, define the Basic Multilingual Dictionary available from the ELRA. Grefenstette (1998) has done a corpus coverage evaluation for the Basic Multilingual Dictionary. The newspapers corpora defined by the US-government-sponsored Text Retrieval Conference (TREC) have been used as a test corpus. The result was that the chance of pulling a random noun out of the different corpora was on average 92%<sup>1</sup>. This statistic is given for the Basic Multilingual Dictionary and, of course, the French Integral Dictionary reaches the highest coverage.

#### 3.2 Semantic links

The links in the semantic network can connect word-senses together, but also classes and concepts. Up to now, more than 100 different kinds of links have been defined. All these links are typed so that a weight can be allocated to each link, given its type. This mechanism allows to very precisely adapt the network to the task: one does not use the same weighting to perform lexical acquisition as to perform word-sense disambiguation. This characteristic makes the network highly adaptive and appropriate to explore some kind of lexical tuning.

This network includes original strategies to measure the semantic proximity between two words. These measures take into account the similarity between words (their common features) but also their differences. The comparison between two words is based on the structure of the graph: the algorithm calculates a score taken into account the common ancestors but also the different ones.

---

<sup>1</sup> This means that for a target English text, one can assume that 92% of the tokens will be in the semantic net.

	A	B	C	D	E	F	G	H
1	SCHEMA	ELT1	CAT1	ELT2	CAT2	SCORE	ETQ	OBJET
2	+	rachat	N	groupe	N	20,787477	entreprise_achetee	\$2
3	+	reprise	N	activités	N	74,256874	entreprise_achetee	\$2
4	+	rachat	N	activité	N	62,731503	entreprise_achetee	\$2
5	+	reprise	N	activités	N	56,257828	entreprise_achetee	\$2
6	-	racheter	V	usine	N	22,668888	entreprise_achetee	\$2
7	-	acquérir	V	usine	N	22,668888	entreprise_achetee	\$2
8	-	racheter	V	c-company	N	44,149246	entreprise_achetee	\$2
9	+	cession	N	société	N	46,118206	entreprise_achetee	\$2

Figure 1: A table of linguistic constraints

We will not detail here the different measures that have been implemented to calculate similarities between words. Please refer to (Dutoit and Poibeau, 2002) for more details.

#### 4 Acquisition of semantically equivalent predicative structures

For IE applications, defining an appropriate set of extraction pattern is crucial. That is why we want to validate the proposed measures to extend an initial set of extraction patterns.

##### 4.1 The acquisition process

The process begins when the end-user provides a predicative linguistic structure to the system along with a representative corpus. The system tries to discover relevant parts of text in the corpus based on the presence of plain words closely related to the ones of the example pattern. A syntactic analysis of the sentence is then done to verify that these plain words correspond to a predicative structure. The method is close to the one of E. Morin et C. Jacquemin (1999), who first locate couples of relevant terms and then try to apply relevant patterns to analyse the nature of their relationship. The detail algorithm is described below:

1. The head noun of the example pattern is compared with the head noun of the candidate pattern using the proximity

measure. This result of the measure must be under a threshold fixed by the end-user.

2. The same condition must be filled by the “expansion” element (the complement of the noun or of the verb of the candidate pattern).
3. The structure must be predicative (either a nominal or a verbal predicate, the algorithm does not make any difference at this level).

The result of this analysis is a table that represent predicative structures equivalent to the initial example pattern. The process uses the corpus and the semantic net as two different complementary knowledge sources:

- The semantic net provides information about lexical semantics and relations between words
- The corpus attests possible expressions and filter irrelevant ones.

We performed some evaluation on different French corpora, given that the semantic net is especially rich for this language. We take the expression *cession de société* (*company transfer*) as an initial pattern. The system then discovered the following expressions, each of them being semantically related to the initial pattern :

reprise des activités

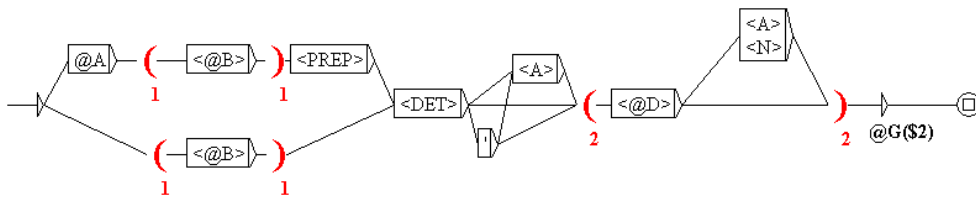


Figure 2: A meta-graph encoding syntactic variations

rachat d'activité  
 acquérir des magasins  
 racheter \*c-company\*  
 cession de \*c-company\*...

This result includes some phase with \*c-company\*: the corpus has been previously preprocessed so that each named entity is replaced by its type. This process normalizes the corpus so that the learning process can achieve better performance.

The result must be manually validated. Some structures are found even if they are irrelevant, due to the activation of irrelevant links. It is the case of the expression *renoncer à se porter acquéreur* (to give up buying sthg), which is not relevant. In this case, there was a spurious link between *to give up* and *company* in the semantic net.

#### 4.2 Dealing with syntactic variations

The previous step extract semantically related predicative structures from a corpus. These structures are found in the corpus in a certain linguistic structure, but we want the system to be able to find this information even if it appears in other kind of linguistic sequences. That is the reason why we associate some meta-graphs with these linguistic structures, so that different transformation can be recognized<sup>2</sup>. This transformation concerns the syntactic level, either on the head (H) or on the expansions (E) of the linguistic structure.

The meta-graphs encode transformations concerning the following structures:

- Subject — verb,
- Verb — direct object,

- Verb — direct object (especially when introduced by the French preposition *à* or *de*),
- Noun — noun complement.

These meta-graphs encode the major part of the linguistic structures we are concern with in the process of IE.

The graph on Figure 2 recognizes the following sequences (in brackets we underline the couple of words previously extracted from the corpus):

Reprise des activités charter... (H: reprise, E: activité)  
 Reprendre les activités charter... (H: reprendre, E: activité)  
 Reprise de l'ensemble des magasins suisse... (H: reprise, E: magasin)  
 Reprendre l'ensemble des magasins suisse... (H: reprendre, E: magasin)  
 Racheter les différentes activités... (H: racheter, E: activité)  
 Rachat des différentes activités... (H: rachat, E: activité)

This kind of graph is not easy to read. It includes at the same time some linguistic tags and some applicability constraints. For example, the first box contains a reference to the @A column in the table of identified structures. This column contains a set of binary constraints, expressed by some signs + or -. The sign + means that the identified pattern is of type verb-direct object: the graph can then be applied to deal with passive structures. In other words, the graph can only be applied in a sign + appears in the @A column of the constraints table. The constraints are removed from the instantiated graph<sup>3</sup>. Even if the resulting graph is normally not visible (the compilation process directly

<sup>2</sup> A meta-graph corresponds to a non-lexicalized graph. A meta-graph is then a kind of abstract grammar (see also the notion of metagrammar in the TAG theory (Candito, 1999))

<sup>3</sup> In other words, an abstract graph is a non-lexicalized graph and an instantiated graph is a lexicalized graph.

produced a graph in a binary format), we can give an equivalent graph.

This mechanism using constraint tables and meta-graph has been implemented in the finite-state toolbox INTEX (Silberztein, 1993). 26 meta-graphs have been defined modelling linguistic variation for the 4 predicative structures defined above. The phenomena mainly concern the insertion of modifiers (with the noun or the verb), verbal transformations (passive) and phrasal structures (relative clauses like ...*Vivendi, qui a racheté Universal... Vivendi, that bought Universal*).

The compilation of the set of meta-graphs produces a graph made of 317 states and 526 relations. These graphs are relatively abstract but the end-user is not intended to directly manipulate them. They generate instantiated graphs, that is to say graphs in which the abstract variables have been replaced linguistic information as modeled in the constraint tables.

This method associates a couple of elements with a set of transformation that covers more examples than the one of the training corpus. This generalization process is close to the one imagined by Morin and Jacquemin (1999) for terminology analysis.

## 5 Evaluation

The evaluation concerned the extraction of information from a French financial corpus, about companies buying other companies. The corpus is made of 300 texts (200 texts for the training corpus, 100 texts for the test corpus).

A system was first manually developed and evaluated. We then tried to perform the same task with automatically developed resources, so that a comparison is possible. At the beginning, the end-user must provide a set of relevant pattern to the acquisition system. We have developed a filtering tool to help the end user focus on relevant portion of text. Due to lack of place, we will not describe this filtering tool, which is very close in its conception to the EXDISCO system developed by R. Yangerber at NYU.

First of all, the corpus is normalized. For example, all the company names are replaced by a variable *\*c-company\** thanks to the named entity recognizer. In the semantic network, *\*c-company\** is introduced as a synonym of company, so that all the sequences with a proper name corresponding to a company could be extracted.

For the slot corresponding to the company that is being bought, 6 seed patterns were given to semantic expansion module. This module acquired from the corpus 25 new validated patterns. Each example pattern generated 4.16 new patterns on average. For example, from the pattern *rachat de \*c-company\** we obtain the following list:

```
reprise de *c-company*
achat de *c-company*
acquérir *c-company*
racheter *c-company*
cession de *c-company*
```

This set of pattern includes nominal phrases (*reprise de \*c-company\**) and verbal phrases (*racheter \*c-company\**). The acquisition process concerns at the same time, the head and the expansion. This technique is very close to the co-training algorithm proposed for this kind of task by E. Riloff and R. Jones (Riloff et Jones, 1999) (Jones *et al.*, 1999).

The proposed patterns must be filtered and validated by the end-user. We estimate that generally 25% of the acquired pattern should be rejected. However, this validation process is very rapid: a few minutes only were necessary to check the 31 proposed patterns and retain 25 of them.

We then compared these results with the ones obtained with the manually elaborated system. The evaluation concerned the two slots that necessitate a syntactic and semantic analysis: the company that is buying another one (slot 1) and the company that is being bought (slot 2). These slots imply nominal phrases, they can be complex and a functional analysis is most of the time necessary (is the nominal phrase the subject or the direct object of the sentence?). An overview of the results is given below (P is for precision, R for recall; P&R is the combined ratio of P and R):

	Slot 1	Slot 2
<b>Human annotators</b>	P: 100 R: 90 ----- P&R : 94.7	P: 100 R: 91.6 ----- P&R : 95.6
<b>INTEX + manual resources</b>	P: 79.6 R: 62.6 ----- P&R : <b>70</b>	P: 93.4 R: 73 ----- P&R : <b>81.9</b>
<b>INTEX + SemTex</b>	P: 65.8 R: 58.7 ----- P&R : <b>62</b>	P: 77 R: 65.3 ----- P&R : <b>70.7</b>

The system running with automatically defined resources is about 10% less efficient than the one with manually defined resources. The decrease of performance may vary in function of the slot (the decrease is less important for the slot 1 than for the slot 2). Two kind of errors are observed:

Certain sequences are not found because a relation between words is missing in the semantic net. This is the case for some idiomatic expressions that were not registered in the network like *tomber dans l'escarcelle de* which means *to acquire*.

Some sequences are extracted by the semantic analysis but do not correspond to a transformation registered in the syntactic variation management module. For example the sequence:

*\*c-company\** renforce son activité communication ethnique en prenant une participation dans *\*c-company\** <sup>4</sup>

is not completely recognized. The pattern (prendre <DET>) participation dans *\*c-company\** correctly identifies the company that is being bought. But the pattern *\*c-company\** (prendre <DET>) participation cannot apply because the subject is too far from the verb.

Lastly, we can mention that some patterns that were not found manually are identified by the automatic procedure. The gain concerning development time is very significant (50 h were necessary to manually define the

<sup>4</sup> *\*c-company\** reinforces its activity in ethnic communication by taking some interest in *\*c-company\**

resources, only 10 h with the semi-automatic process).

Even if the decrease of performance is significant (10%), it can be reduced using more linguistic knowledge. For example, we know that nominalizations are not correctly handled by the system at the moment. Some more information could be used from the semantic network (that also includes morphological and syntactic information) to enhance the performances of the overall system.

Experiments have been made on different corpora and on different MUC-like tasks. They have all proved the efficiency of the strategy described in this paper. Moreover, it is possible to adapt the system so that it has a better precision, or a better recall, given user needs (Poibeau, 2001). For example, people working on large genomic textual databases are facing a huge amount of redundant information. They generally want some very precise information to be extracted. On the other hand, human operators monitoring critical situation generally want to be able to have access to all the available information. Our system is versatile and could be easily adapted to these different contexts.

## 6 Conclusion

In this paper, we have shown an efficient algorithm to semi-automatically acquire extraction patterns from a semantic network and a corpus. Even if the performance decrease when the resource are automatically defined, the gain in development time is sufficiently significant to ensure the usability of the approach.

## 7 References

- Appelt D.E, Hobbs J., Bear J., Israel D., Kameyana M. and Tyson M. (1993) FASTUS: a finite-state processor for information extraction from real-world text. Proceedings of IJCAI'93, Chambéry, France, pp. 1172—1178.
- Bagga A., Chai J.Y. et Biermann A. The Role of WORDNET in the Creation of a Trainable Message Understanding System. In Proceedings of the 14<sup>th</sup> National Conference on Artificial Intelligence and

- the Ninth Conference on the Innovative Applications of Artificial Intelligence (AAAI/IAAI'97), Rhode Island, 1997, pp. 941–948.
- Bikel D., Miller S., Schwartz R. and Weischedel R. (1997) Nymble: a high performance learning name-finder. Proceeding of the fifth Conference on Applied Language Processing, Washington, USA.
- Candito, M.-H. Organisation modulaire et paramétrable de grammaires électroniques lexicalisées. PhD Thesis, University Paris 7, 1999.
- Ciravegna F. Adaptive Information Extraction from Text by Rule Induction and Generalisation. In Proceedings of the 17<sup>th</sup> International Joint Conference on Artificial Intelligence (IJCAI'2001), Seattle, 2001, pp. 1251–1256.
- Dutoit D. and Poibeau T. (2002) Inferring knowledge from a large semantic network. In Proceedings of COLING'2002, Taïpei.
- Fellbaum C. (1998) WordNet : An Electronic Lexical Database, edited by Fellbaum, M.I.T. press.
- Freitag D. (1998) Machine learning for Information Extraction in Informal Domains, Thesis, Carnegie Mellon University, USA.
- Grefenstette G. (1998) Evaluating the adequacy of a multilingual transfer dictionary for the Cross Language Information Retrieval, LREC 1998.
- Jones R., McCallum A., Nigam K. and Riloff E. (1999) Bootstrapping for Text Learning Tasks. Proceedings of the IJCAI'99 Workshop on Text Mining: Foundations, Techniques and Applications, Stockholm, 1999, pp. 52–63.
- Morin E. and Jacquemin C. (1999) Projecting corpus-based semantic links on a thesaurus. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL'99), Maryland, 1999, pp. 389–396.
- Muslea I. (1999) Extraction patterns for Information Extraction tasks: a survey, AAAI'99 (available at the following URL: <http://www.isi.edu/~muslea/RISE/ML4IE/>)
- Pazienza M.T, ed. (1997) Information extraction. Springer Verlag (Lecture Notes in computer Science), Heidelberg, Germany.
- Poibeau T. (2001) – « Deriving a multi-domain information extraction system from a rough ontology. Proceeding of the 17<sup>th</sup> International Conference on Artificial Intelligence (IJCAI'2001), Seattle, 2001, pp. 1264–1270.
- Riloff E. (1993) Automatically constructing a dictionary for formation extraction tasks, AAAI'93, Stanford, USA, pp. 811–816.
- Riloff E. (1995) Little Words Can Make a Big Difference for Text Classification , Proceedings of the SIGIR'95, Seattle, USA, pp. 130–136.
- Riloff E. et Jones R. (1999) Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. Proceedings of the 16th National Conference on Artificial Intelligence (AAAI'99), Orlando, 1999, pp. 474–479.
- Silberztein M. (1993) Dictionnaires électroniques et analyse automatique des textes, Masson, Paris, France.
- Soderland S., Fisher D., Aseltine J. and Lenhart W. (1995) Crystal: inducing a conceptual dictionary, Proceedings of IJCAI'95, Montréal, Canada, pp. 1314–1319.
- Yangarber R. (2000) Scenario Customization for Information Extraction. PhD Thesis, New York University.