Automatic extraction of paraphrastic phrases from small size corpora

Thierry Poibeau, Dominique Dutoit

Automatic Extraction of Paraphrastic Phrases from Small Size Corpora

Thierry Poibeau[*] and Dominique Dutoit[**]
[*] *Laboratoire d'Informatique de Paris-Nord*
[**] ***Memodata et CRISCO***

Introduction

The recognition of paraphrases is of paramount importance for natural language processing. It is well known that one of the main difficulties in the field is to be able to identify precise pieces of information, despite the fact that the same information can be expressed using a wide range of different constructions. For example, a system should be able to identify that all the following sentences express the same information:

Jack Ruby killed Lee Harvey Oswald,
Lee Harvey Oswald was killed by Jack Ruby
Lee Harvey Oswald was assassinated by Jack Ruby
The murder of Lee Harvey Oswald by Jack Ruby happened in the 60's.

This example is relatively simple but it illustrates the fact that variation requires complex rules, so that it is then possible for a system to automatically compute equivalencies between sentences (*X kills Y* ≅ *X assassinates Y* ≅ *the murder of Y by X*…). These rules include lexical, syntactic and semantic knowledge.

Recent research in NLP has promoted a now widely-accepted shallow-based analysis framework that has proven to be efficient for a number of tasks, including information extraction and question answering. However, this approach often leads to over-simplified solutions to complex problems. For example, surface methods that do not use syntactic knowledge may fail in examples such as: *Lee Harvey Oswald, the gunman who assassinated President John F. Kennedy, was later shot and killed by Jack Ruby* (example taken from J. Lin and B. Katz, 2003). In this case, it is essential to keep track of the argument structure of the verb, to be able to infer that it is *Jack Ruby* and not *John Kennedy* who is the murderer of *Lee Harvey Oswald*. A wrong result would be obtained considering too shallow analysis techniques or too simplistic heuristics, based for example of the proximity between two person names in the sentence (since *Oswald* in closer to *Kennedy* than *Ruby* in the sentence).

Several studies (see, for example, D. Lin and P. Pantel, 2002 and some other references in the next section) have recently proposed various approaches based on the redundancy of the web to acquire extraction patterns and semantic structures. However, most of the time these methods cannot be applied to small-size corpora (< 100,000 words), since these corpora are

not regular or redundant enough. Moreover, existing structured knowledge contained in dictionaries, thesauri or semantic networks can boost the learning process by providing clear intuition over text units.

In this paper, we propose a knowledge rich approach to paraphrase acquisition. We will firstly describe some related work for the acquisition of knowledge, especially paraphrases, from texts. We then describe how semantic similarity between words can be inferred from a large semantic network. We present an 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 adaptation process (this process is also known as *lexical tuning*, see Y. Wilks and R. Catizone, 2002), so that one can generate paraphrases of an original pattern from the information contained in the semantic network. In the last section, we propose an evaluation and some perspectives.

1 Related work

As it has been shown in the introduction, both lexical and syntactic knowledge is necessary for paraphrase detection. Semantic classes (related words occurring in similar contexts) are thus highly relevant since they are parts of paraphrase. This section presents related works for the acquisition of semantic classes and extraction patterns from texts.

1.1 Acquisition of semantic classes

Several studies have recently proposed measures to calculate the semantic proximity between words. Different measures have been described, which are not easy to evaluate (see D. Lin and P. Pantel (2002) for some evaluation proposals). The methods proposed so far are automatic or manual and generally imply the evaluation of word clusters in different contexts (a word cluster is close to another one if the words it contains are interchangeable in some linguistic contexts).

A. Budanitsky and G. Hirst (2001) present the evaluation of 5 similarity measures based on the structure of Wordnet. All the algorithms they examine are based on the hypernym-hyponym relation, which structures the classification of clusters inside Wordnet (the *synsets*). They sometimes obtain unclear conclusions about the reason of the performances of the different algorithms (for example, comparing J. Jiang and D. Conrath's measure (1997) with D. Lin's one (1998): "It remains unclear, however, just why it performed so much better than Lin's measure, which is but a different arithmetic combinations of the same terms").

Moreover, the authors stress the fact that the use of the sole hyponym relation is insufficient to capture the complexity of meaning: "Nonetheless, it remains a strong intuition that hyponymy is only one part of semantic relatedness; meronymy, such as *wheel–car*, is most definitely an indicator of semantic relatedness, and, *a fortiori*, semantic relatedness can arise from little more than common or stereotypical associations or statistical co-occurrence in real life (for example, *penguin–Antarctica; birthday–candle; sleep–pajama*s)". In this paper, we propose a method based on a rich semantic network containing different kinds of semantic relations, to overcome some of the limitations of the previous approaches.

1.2 Acquisition of extraction patterns

IE is known to have established a now widely accepted linguistic architecture based on cascading automata (*i.e.* automata than can be applied recursively) and domain-specific

knowledge (D. Appelt *et al.*, 1993). However, several studies have outlined the problem of the definition of the resources. For example, E. Riloff (1995) says that about 1,500 hours are necessary to define the resources for an information extraction system. Most of these resources are in fact variants of extraction patterns, which have to be manually described. To address this problem, a recent research effort focused on using machine learning throughout the IE process (I. Muslea, 1999). A first trend was to directly apply machine learning methods to replace IE components. Statistical methods have been successfully applied to the named-entity task. For example, D. Bikel *et al.* (1997) learn 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. I. Muslea (1999) gives an extensive description of the different approaches of that problem. Autoslog (E. Riloff, 1995) was one of the very first systems using a simple form of learning to build a dictionary of extraction patterns. F. Ciravegna (2001) demonstrates the interest of independent acquisition of left and right boundaries of extraction patterns during the learning phase. In general, the left part of a pattern is easier to acquire than the right part but some heuristics can be applied to infer the right boundary from the left one. The same method can be applied for argument acquisition: each argument can be acquired independently from the others since the argument structure of a predicate in context is rarely complete.

M. Collins and Y. Singer (1999) demonstrate how two classifiers operating on disjoint features sets recognize named entities with very little supervision. The method is interesting in that the analyst only needs to provide some seed examples to the system in order to learn relevant information. However, these classifiers must be made interactive in order not to diverge from the expected result: otherwise, each error is transmitted and amplified by subsequent processing stages. Contrary to this approach, partially reproduced by F. Duclaye *et al.* (2003) for paraphrase learning, we prefer a slightly supervised method with clear interaction steps with the analyst during the acquisition process, to ensure the solution is converging (*i.e.* control the fact that the precision ratio does not decrease abruptly due to a wrong inference).

Since the 2000s, several studies have proposed fully automatic approaches to relation extraction, especially for information extraction and question answering, see (D. Lin and P. Pantel, 2002; D. Ravichandran and E. Hovy, 2002; N. Català *et al.* 2003; A. Agichtein and L. Gravano, 2000; I. Szpektor *et al.*, 2004) among many others. All these approaches are based on seed elements (a list of verbs for I. Szpektor *et al.*, 2004; a list of patterns for D. Lin and P. Pantel, 2001): the idea is then to acquire information on the context by generalization algorithms. The main idea is Z. Harris' hypothesis (1968): according to the pioneering work from R. Grishman *et al.* (1986): "Zellig Harris, one of the first linguists to study language use in restricted domains, defined sublanguages in terms of one particular constraint: the constraint on what words can co-occur within a particular syntactic pattern, such as a subject-verb-object structure". Even if this approach has proved to be successful, it is easy to show its limitations. Rare events are not captured by automatic methods based on redundancy detection and, consequently, this method cannot be applied successfully to small corpora, which are not regular enough and contain rare constructions. We thus propose to extend the method with other resources, especially a semantic network, to overcome data sparseness.

2 Overview of the approach

Argument structure acquisition is a complex task since the argument structure is rarely complete in corpora. To overcome this problem, we propose an acquisition process in which all the arguments are acquired separately. Figure 1 presents an outline of the overall

paraphrase acquisition strategy. The process is made of automatic steps and manual validation stages. The process is weakly supervised since the analyst only has to provide one example to the system. However, we observed that the quality of the acquisition process highly depends from this seed example. It is this necessary to re-iterate the process several times, so that we are sure to obtain an accurate coverage of the corpus.

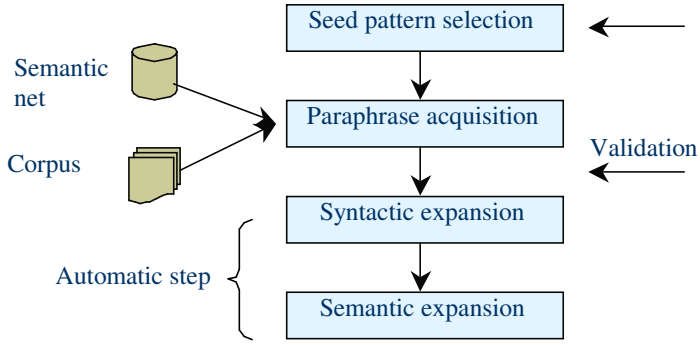


Figure 1: overview of the approach

From the seed pattern, a set of paraphrases is automatically acquired, using similarity measures between words and a shallow syntactic analysis of the found patterns, in order to ensure they describe a predicative sequence.

The overall approach is based on a large knowledge base used to avoid data sparseness. This knowledge base mainly consists in a semantic network with typed relations between words (see next section).

3 The semantic network

The semantic network used in this experiment is a multilingual network providing information for 5 European languages. In this section, we describe the network and then give some details about its overall structure (see D. Dutoit, 2000, for more details).

3.1 Overall organisation

The semantic network we used is called *The Integral Dictionary* (D. Dutoit, 2000 and 2004). This database is basically structured as a merging of three semantic models available for five languages. The maximal coverage is obtained for the French language, with 185,000 word-meanings encoded in the database. English appears like the second language in terms of coverage with 79,000 word-meanings. Three additional languages (Spanish, Italian and German) include about 39,500 senses.

These multilingual dictionaries, with universal identifiers to ensure the translation, define the Basic Multilingual Dictionary available from the ELRA. G. Grefenstette (1998) has done an evaluation of the coverage of the Basic Multilingual Dictionary: a corpus developed from newspapers (the corpus defined for the Text Retrieval Evaluation Conference, TREC, 2000) has been used as a test corpus and the experiment consisted in checking the coverage of the dictionary. The result was on average 92% (this score is given for the whole Multilingual Dictionary but, of course, among the multilingual dictionaries, the French Integral Dictionary reaches the highest coverage).

WordNet is a semantic network whose design is inspired by psycholinguistic theories of human lexical memory. English nouns, verbs, adjectives and adverbs are organized into

synonym sets, each representing one underlying lexical concept (C. Fellbaum, 1998). The Integral Dictionary is richer than the French Wordnet: it has got a larger number of links and is based on a componential lexical analysis (decomposition of words into basic semantic units). Therefore, words are highly interconnected in the semantic network.

Relations between terms in the Integral Dictionary include classical ones like synonymy, hyponymy (*specific*), hypernymy (*generic*), meronymy (*part-of*), *etc*. Several dozens of functions, inspired from Mel'cuk theory, make it possible to describe collocations and unpredictable links between words. For example `Magn(pleuvoir)="à verse"` is the `Magn` function (expressing intensity) applied to the French verb *pleuvoir* (*to rain*), because the meaning of the French expression *pleuvoir à verse* (*to pour down*) cannot be directly inferred from the words it contains (Mel'cuk, 1996). These functions are not used in this work, but they will be used in future evolutions since idioms have to be taken into account.

3.2 Link weighting

The network is made of more than 220,000 relations between nodes (i.e. between word senses or between concepts). Each link is coded according to a type hierarchy. The basic structure is of course sketched by the generic/specific relation but several other relations are also encoded like *is-synonym-of*, *is-part-of*, *see-also*, etc. Each relation is then refined with sub-relations. For example, *antonyms*, *acronyms*, *abbreviations*, *etc*. are considered as specific cases of *synonymy*. In the following experiments, we do not take into account these subtypes, we only refer to the first level of the hierarchy.

All the links in the semantic network are typed so that a weight can be allocated to each link, given its type. This mechanism enables us to tune with precision 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 convenient to explore the ramifications of lexical tuning.

For the acquisition of extraction patterns, several relations are favoured whereas some others are discarded from the semantic network (they are not materially discarded but their respective weight is set to 0). These relations are the following:

1. The synonymy relation is the most relevant relation for the application.
2. The hyponym (*specific*) and hypernym (*generic*) relation cannot be used in the same way. On the one hand, the generic relation generates a lot of noise (because it captures too general terms like *trading* (*commerce*), *commercial activity* (*activité comemrciale*)…). On the other hand, we observed that the hyponym relation (*specific*) generates a lot of interesting terms and is nearly as productive as the synonymy relation.
3. The part-of relation (*meronyms*) is very close to the hyponym relation and it is sometimes hard to distinguish among these two relations. This relation is highly relevant for our application.
4. The see-also relation links words like *goods* (*marchandise*) or *trade* (*commerce*) to *purchase* (*achat*). It is useful for tasks such as word sense disambiguation but is too general for the extraction task. It is thus discarded from the network.
5. Verbal relations were penalized in the original network, since most NLP applications are based on technical terms and nominal phrase detection. However, for our application, verbal phrases are rather important. The relation is thus favoured.

Weights for the different relations were stored in the database. They had to be manually changed to reflect the previous observations. This is done through a "try and error" strategy. In the end, the relations between terms are homogeneous, independently from their relative

part-of-speech. Weights also depend on the length of the path between two nodes. For example, if A is a hypernym of B and B and hyperonym of C, the link between A and B must be stronger than the link between A and C.

3.3 Similarity measures

We propose an original way to measure the semantic proximity between two word senses (see D. Dutoit and T. Poibeau, 2002 for more details). This measure takes 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 taking into account the common ancestors but also the different ones. The notion of "nearest common ancestor" is classical in graph theory. In (Dutoit and Poibeau, 2002), we extend this notion to distinguish between "symmetric nearest common ancestor" (direct common ancestor for both nodes) and "asymmetric nearest common ancestor" (common ancestor, indirect at least for one node).

Definition: Distance between two nodes in a graph
We note `d` the distance between two nodes `A` and `B` in a graph. This distance is equivalent to the number of arcs between two nodes `A` and `B`. We have `d(A, B) = d(B,A)`.

Let's say :
`h(A)` = the set of ancestors of `A` .
`c(A)` = the set of arcs between a node `A` and the graph's root.

Definition: Nearest common ancestors (NCA)
The nearest common ancestors between two words `A` and `B` are the set of nodes that are daughters of `c(A) ∩ c(B)` and that are not ancestors in `c(A) ∩ c(B)`.

We then propose a measure to calculate the similarity between two words. The measure is called *activation* and only takes into account the common features between two nodes in the graph. An equal weight is attributed to each NCA. This weight corresponds to the minimal distance between the NCA and each of the two concerned nodes.

Definition: activation ($d_{\curlywedge}$)

The *activation* measure $d_{\curlywedge}$ is equal to the mean of the weight of each NCA calculated from `A` and `B` :

$$d_{\curlywedge}(A, B) = \frac{1}{n}\sum_{i=1}^{n}(d(A,NCAi)+d(B,NCAi))$$

The activation measure has the following properties:
- $d_{\curlywedge}(A, A) = 0$, because `A` is the unique NCA of `A` $\curlywedge$ `A`.
- $d_{\curlywedge}(A, B) = d_{\curlywedge}(B, A)$ (symmetry)
- $d_{\curlywedge}(A, B) + d_{\curlywedge}(B, C) >= d_{\curlywedge}(A, C)$ (euclidianity)

The set of NCA takes into account the common features between two nodes `A` et `B`. We then need another measure to take into account their differences; we must first define the notion of asymmetric nearest common ancestor.

---

[12] This measure allows comparing two sets of words, or two sentences. For a sentence, it is first necessary to delete empty words, to obtain a set of full words.

Definition: Asymmetric nearest common ancestor (ANCA)
The asymmetric nearest common ancestors from a node A to a node B is contained into the set of ancestors of c(B) ∩ c(A) which have a direct node belonging to h(A) but not to h(B).

It is now possible to measure the distance between two words from their differences. A weight is allocated to each link going from node N_i, asymmetric nearest common ancestor, to A and B. The weight is equal to the minimal length of the path going from A to N_i and from B to N_i.

Definition: proximity ($d_\perp$)
The proximity measure takes into account the common ancestors but also the differences between two elements A and B and is defined by the following function:

$$d_\perp(A,B) = d_\wedge(A,B) + \frac{1}{n}\sum_{i=1}^{n}(d(A,ANCA_i)+d(B,ANCA_i))$$

Because the set of ANCA from a node A to a node B is not the same as the one from a node B to a node A, the proximity measure has the following properties:

- $d_\perp(A, A) = 0$, because $ANCA(A, A) = \varnothing$.
- $d_\perp(A, B) \neq d_\perp(B, A)$ if the set of ANCA is not empty (anti-symmetry)
- $d_\perp(A, B) + d_\perp(B, C) >= d_\perp(A, C)$ (euclidianity)

The proximity measure is dependent from the structure of the network. However, this measure is relative: if the semantic network evolves, all the proximity measures between nodes could be changed. However, the relations between nodes stay relatively stable since the network is highly inter-connected, so minor revisions of it do not change the overall network connectivity.

Since it is not symmetric, the proximity measure can discriminate between two words, whereas the activation measure cannot. Therefore, the componential analysis of the semantic network is able to reflect some weak semantic differences between words.

4 Paraphrase acquisition

Our aim is to acquire paraphrases from the corpus, using the measures defined previously. The process begins as the end-user provides a predicative structure (a predicate with its arguments) to the system along with a representative corpus (preferably, a corpus from the same source than the one that will be further analyzed by the system). The system tries to discover relevant parts of text in the corpus based on the presence of plain words closely related to the initial predicative structure (the *seed pattern*). A syntactic analysis of the sentence is then done to verify that these plain words correspond to a paraphrastic structure. The method is close to the one of E. Morin and C. Jacquemin (1999), who first try to locate couples of relevant terms and then apply relevant patterns to analyse the nature of their relationship. However, E. Morin and C. Jacquemin only focus on term variations whereas we are interested in predicative structures, being either verbal or nominal. The syntactic variations we have to deal with are then different and, for a part, more complex than the ones examined by E. Morin and C. Jacquemin.

The detailed algorithm is described below:

1. The head word (a noun or a verb) of the example pattern is compared with the head word of the candidate pattern using the proximity measure introduced in 4.3. The result of the proximity measure must be under a threshold fixed by the end-user.
2. The same condition must be filled by the "expansion" element (possessive phrase or verb complement in 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 following schema (Figure 2) resumes the acquisition process.

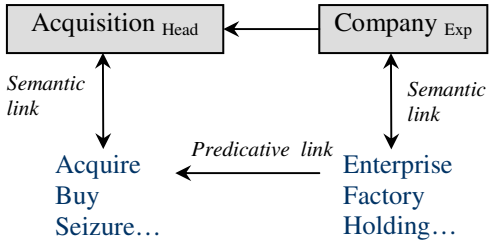


Figure 2: paraphrase acquisition

Finally, this process is formalized throughout the algorithm 1. Note that the predicative form is acquired together with its arguments, as in a co-training process.

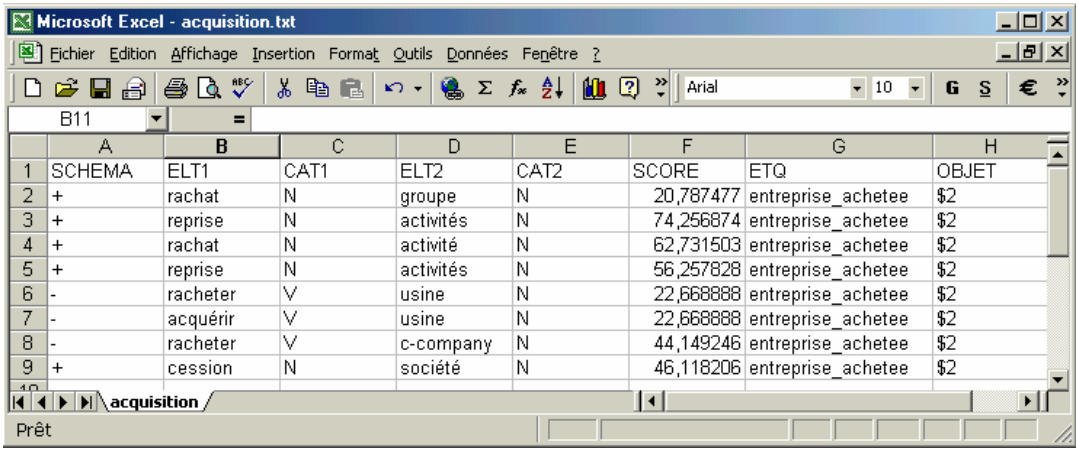
```
P ← pattern to be found
S ← Sentence to be analyzed
C ← Phrases(S)
W ← Plain_words(S)
Result ← empty list
head ← Head word of the pattern P
exp ← Expansion word of the pattern P
Threshold ← threshold fixed by the analyst
For every word wi from W do
 Prox1 = d'⊥(head, wi)
 If (Prox1 <= Threshold) then
   wi+1 ← Next element from W (if end of sentence then exit)
  Prox2 = d'⊥(exp, wi+1)
  If (Prox2 <= Threshold) then
    If there is c ∈ C so that (wi ∈ c) and (wi+1 ∈ c) then
       Result ← Add (wi, wi+1)
    End_if
   End_if
  End_if
End_for
```

Algorithm 1 : Paraphrastic phrases acquisition

The result of this analysis is a table representing predicative structures, which are semantically equivalent to the initial example pattern (Figure 3). 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 is used to attest possible expressions and filter irrelevant ones.

We performed an evaluation on different French corpora, given that the semantic net is especially rich for this language. We took the expression *cession de société* (*company*



	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 3: the linguistic constraint table

transfer) as an initial pattern. The system then discovered the following expressions, each of them being semantic paraphrases of the initial seed pattern:

```
reprise des activités      (trade-in of activities)
rachat d'activité        (repurchase of activities)
acquérir des magasins      (acquire shops)
racheter *c-company*       (to buy another *c-company*)
cession de *c-company*…    (transfer of *c-company*)
```

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.

4.1 Dealing with syntactic variations

The previous step extracts semantically related predicative structures from a corpus. These structures are found in the corpus through various linguistic constructions and we want the system to be able to find this information despite the difference of constructions. A convenient way of modelling syntactic relations between items is to use automata; we chose to use "meta-graphs" that is to say empty automata representing a set of variations (one could say a "family resemblance") from a basic construction. For example, if a verb is transitive, the passive construction can be applied most of the time, quite independently from the specific lexical item (Silberztein, 1999).

This strategy is based on Z. Harris' theory of sublanguages (1991). These transformations concern the syntactic level, either on the head (H) or on the expansion part (E) of the linguistic structure. Information captured from the corpus gives predicate-arguments with alternation information. The meta-graphs encode transformations concerning the following structures:

- Subject — verb,
- Verb — direct object,
- Verb — indirect object (especially when introduced by the French preposition *à* or *de*),
- Noun — possessive phrase.

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

The graph on Figure 4 recognizes the following sequences (in brackets we underline the couple of words previously extracted from the corpus):

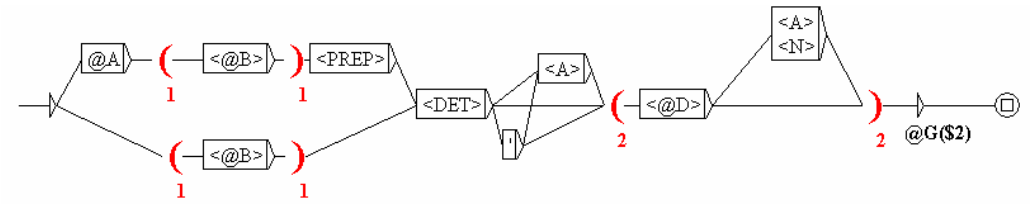


Figure 4: a syntactic meta-graph

```
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 (Figure 4) contains a reference to the `@A` column in the table of identified structures (Figure 3). 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 if a sign + appears in the `@A` column of the constraints table. The constraints are removed from the instantiated graph. Even if the resulting graph is normally not visible (the compilation process directly produced a graph in a binary format), we give an image of a part of that graph on Figure 5.

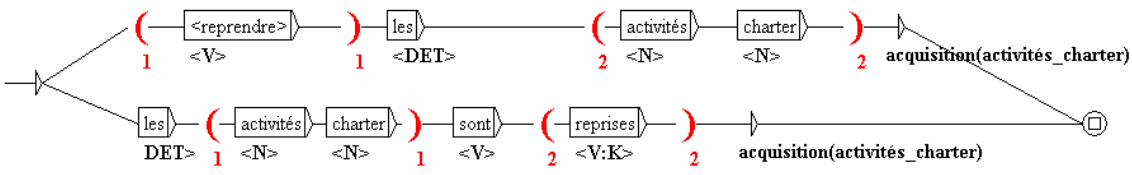


Figure 1 : An instantiated graph automatically generated from a meta-graph and a constraint table.

This mechanism using constraint tables and meta-graph has been implemented in the INTEX/UNITEX finite-state toolbox (M. Silberztein, 1993; S. Paumier, 2002). Twenty six meta-graphs have been defined, modelling linguistic variation for the four 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 new 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 modelled in the constraint tables. This method associates a couple of elements with a set of transformation that covers more examples than the ones of the training corpus. This generalization process is close to the one imagined by E. Morin and C. Jacquemin (1999) for terminology analysis but, as already said, we cover sequences that are not only nominal ones.

5 Experiment and evaluation

In this section we propose an evaluation of the use of the semantic network and of the measures that have been implemented through a set of NLP applications related to information filtering. To help the analyst focus on relevant information in texts, it is necessary to first provide filtering tools, based on a "profile" defined by the end-user himself, that describes his research interests (C. van Rijsbergen, 1979; E. Voorhees, 1999). We present two experiments, in which we use the semantic network as a knowledge base for guiding the filtering and the acquisition of extraction patterns.

5.1 Data

The evaluation concerned the extraction of information from a French financial corpus, about companies buying other companies. The topic was the same as in the MUC-5 conference (1993). The corpus is made of 500 texts (the overall corpus is about 65,000 words) from the financial website *FirstInvest* (now *Edubourse.fr*).

As we have seen previously, the task is twofold: 1) filter relevant sentences from texts and 2) extract information from these sentences. The texts have been annotated by two human annotators. These annotators were asked to manually perform the two tasks (sentence extraction and information extraction) over a corpus of 500 texts. Inter-annotator agreement was high, achieving more than 95% agreement on both tasks.

We used the 500 texts to evaluate the filtering task, which does not require any training corpus. For the evaluation task, 300 texts were used for the training corpus and 200 texts for the test corpus.

5.2 Information filtering

The first step consists in retrieving only parts of the text that are interesting for the task. We thus developed a filtering strategy, according to a profile. A profile is a set of words, describing the user's domain of interest. Unfortunately the measures we have described so far are only concerned with one single word, not with a set of words (see section 4.3).

We then need to slightly modify the activation measure, so that it accepts to compare two sets of words, and not only two words[2]. We propose to aggregate the set of nodes in the graphs corresponding to the set of words in the profile. This node has the following properties:

$$h(M)=\bigcup_{i=1}^{n} h(m_i)$$

$$c(M)=\bigcup_{i=1}^{n} c(m_i)$$

where `h(M)` is the set of ancestors of `M` and `c(M)` the set of links between `M` and the root of the graph. It is then possible to compare two sets of words, and not only two words.

In an IE task, one wants to filter texts to focus on sentences that are of possible interest for the extraction process (sentences that could allow filling a given slot). We then need a very precise filtering process performing at the sentence level[3]. We used the activation measure for the filtering task. A sentence is kept if the activation score between the filtering profile and the sentence is above a certain threshold (empirically defined by the end-user, using a try and error strategy). A filtering profile is a set of words in relation with the domain or the slot to be filled, defined by the end-user.

The filtering profile was composed of the following words: `rachat`, `cession`, `enterprise` (`buy`, `purchase`, `company`). The corpus has been manually processed to identify relevant sentences (the reference corpus). We then compare the result of the filtering task with the reference corpus.

In the different experiments we made, we modified different parameters such as the filtering threshold (the percentage of sentences to be kept). We obtained the following results:

	10%	20%	30%	40%	50%
Precision	.72	.54	.41	.33	.28
Recall	.43	.64	.75	.81	.85

We also tried to normalize the corpus that is to say to replace entities by their type, to improve the filtering process. We see about 5% increase if we only take 10% of the corpus.

	10%	20%	30%	40%	50%
Precision	.75	.56	.43	.34	.29
Recall	.49	.71	.82	.89	.94

We notice that we obtain, from 10% of the corpus, a 75% precision ratio (3 sentences out of 4 are relevant) and nearly a 50% recall ratio. The main interest of this process is to help the end-user directly focus on relevant pieces of text. This strategy is very close to the EXDISCO system developed by R. Yangarber at NYU (2000), even if the algorithms we use are different. Our strategy is based on the semantic distance described in section 4.3. which is based on a knowledge-rich semantic net encoding a large variety of semantic relationships between sets of words, including meronymy (*part-of*) and stereotypical associations. This approach is thus convenient for small size corpora.

5.3 The Information Extraction process

In order to evaluate the extraction process, we first manually developed an information extraction system[4] and evaluated its performances. We then tried to perform the same task with semi-automatically developed resources, so that a comparison is possible. The corpus is firstly normalized: for example, all the company names are replaced by a variable **`*c-`**

[3] This is original since most of the systems so far are concerned with texts filtering, not sentence filtering.

[4] More exactly, it is the resource of the information extraction system that was manually developed.

company* thanks to the named entity recogniser[5]. 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 can be extracted.

For the slot corresponding to the company that is being bought, 6 seed example patterns were given to the 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 paraphrastic patterns 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. The simultaneous acquisition of different semantic classes can also be found in the co-training algorithm proposed for this kind of task by E. Riloff and R. Jones (1999).

The proposed patterns must be filtered and validated by the end-user. The idea is that experts in the field should be able to develop there own set of resources; therefore we assume that the end-user has some knowledge of the domain. He can say if a document is relevant or not. The experiments have been done with three experts in scientific and economic intelligence from the private domain (industry). These people are scientists, they have specific needs and, if a training support is provided, they feel, after a while, confident enough to deal with this kind of applications. They need to be able to develop their own simple extraction system in a few hours. The developing phase is also for them a means to dynamically explore the corpus.

After several experiments, we observed 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 for the `arg1` slot (80% of the proposed patterns are relevant).

We then compared these results with the ones obtained with the manually elaborated system. The evaluation concerned the three slots that necessitate a syntactic and semantic analysis: the company that is buying another one (`arg1`) the company that is being bought (`arg2`), the company that sells (`arg3`). These slots imply nominal phrases, they can be complex and a functional analysis is most of the time necessary (is the nominal phrase subject or direct object of the sentence?).

We thus chose to perform an operational evaluation: what is then evaluated is the ability of a given phrase or pattern to fill a given slot (cf. the notion of textual entailment between different text variants that express the same relation (I. Dagan and O. Glickman, 2004)). This kind of evaluation avoids, as far as possible, the bias of human judgment on possibly ambiguous expressions.

An overview of the results is given below (P refers to precision, R to recall, F to the harmonic mean between P and R):

---

[5] Named entity recognizers are now available for various languages, with a lot of variations in their performances (Poibeau, 2003b). Machine learning techniques make it possible to quickly develop such a tool from annotated corpora. However, results are still far from perfect and machine learning based methods are often difficult to adapt and correct. Most industrial tools are still based on finite state transducers that can be easily tuned to a new task or a new domain.

	Arg 1	Arg 2	Arg 3
Human annotators	P: 100 R: 90	P: 100 R: 91.6	P: 99 R: 92
	F: **94.7**	F: **95.6**	F: **94.2**
Automatically acquired resources	P: 79.6 R: 62.6	P: 93.4 R: 73	P: 88.4 R: 70
	F: **70**	F: **81.9**	F: **77**

We observed that 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 `arg2` than for `arg1` or `arg3`). Two kinds of errors are observed: certain sequences are not found because a relation between words is missing in the semantic net. Some sequences are extracted by the semantic analysis but do not correspond to a transformation registered in the syntactic variation management module. A proper treatment of nominalizations would enhance recall; this was not done at this stage of our work.

These results cannot be integrated directly in a database since the error rate is important (F-measure between 70 and 80). Note, however, that these results are comparable with those from other systems performing the same kind of task. Moreover, it has been proved that they provide a useful aid for analysts, even in an industrial environment (Poibeau, 2003). Relevant ergonomic strategies have to be used to overcome the difficulties and errors generated by automatic approaches: for example, relevant pieces of information are highlighted inside the text and not extracted from the text, so that the analyst can quickly check the context; hyperlinks allow to navigate inside the corpus which makes it easy to discover changes and evolutions, etc.

Conclusion

In this paper, we have shown an efficient algorithm to semi-automatically acquire paraphrastic phrases from a semantic net and a corpus. We have shown that this approach is highly relevant in the framework of IE systems, especially for small size corpora, where a linguistic resource compensate data sparseness. Even if the performance decreases when the resources are automatically defined, the gain in terms of development time is sufficiently significant to ensure the usability of the method.

This module was first design in an industrial environment and has been used by expert users, so that they are able to define by themselves resources in function of their own interests and applications. These experts are scientists who need to quickly access small and medium size corpora in order for them to write synthetic documents and reports. They need to be trained during an initial learning phase before being really independent and define their own resources without any help. However, this is not a major problem since it has been proved by experiments with real users that this software fits their needs.

Some more automatic methods appear nowadays and could learn a part of required knowledge from less annotated data than in the past few years. In the future, it would be interesting to mix these different techniques, in order to limit the amount of work currently devoted to the end-user.

References

Agichtein Eugene and Gravano Luis. 2000. Snowball: Extracting relations from large plain text collections. *International Conference on Digital Libraries* (ICDL). Kyoto.

Appelt Douglas, Hobbs Jerry, Bear John, Israel David, Kameyama Megumi and Tyson Mabry. 1993. FASTUS: a finite-state processor for information extraction from real-world text. *Joint Conference on Artificial Intelligence* (IJCAI). Chambéry. pp. 1172–1178.

Bagga Amit, Chai Joyce and Biermann Alan. 1997. The Role of WORDNET in the Creation of a Trainable Message Understanding System. *National Conference on Artificial Intelligence and the Ninth Conference on the Innovative Applications of Artificial Intelligence* (AAAI/IAAI). Rhode Island. pp. 941–948.

Bikel Daniel, Miller Scott, Schwartz Richard and Weischedel Ralph. 1997. Nymble: a high performance learning name-finder. *Applied Natural Language Processing Conference* (ANLP). Washington.

Budanitsky Alexander and Hirst Graeme. 2001. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. *NAACL Workshop "WordNet and Other Lexical Resource"*. Pittsburgh.

Català Neus, Castell Nuria and Martín Mario. 2003. A portable method for acquiring information extraction patterns without annotated corpora. *Natural Language Engineering*. n° 9(2), pp. 151–179.

Ciravegna Fabio. 2001. Adaptive Information Extraction from Text by Rule Induction and Generalisation. *International Joint Conference on Artificial Intelligence* (IJCAI). Seattle. pp. 1251–1256.

Collins Michael and Singer Yoram. 1999. Unsupervised models for named entity classification. *Empirical Methods in Natural Language Processing workshop* (EMNLP). College Park. pp. 100–110.

Dagan Ido and Glickman Oren. 2004. Probabilistic Textual Entailment: Generic Applied Modelling of Language Variability. *Workshop Learning Methods for Text Understanding and Mining*. Grenoble.

Duclaye Florence, Yvon François and Collin Olivier. 2003. Learning paraphrases to improve a question answering system. *EACL Workshop "NLP for Question Answering"*. Budapest.

Dutoit Dominique. 2000. A text->meaning->text dictionary and process". *Language resource and evaluation conference* (LREC). Athens.

Dutoit Dominique and Poibeau Thierry. 2002. Combining knowledge sources for resource acquisition. *Proceeding of the Computational Linguistics Conference* (COLING), Taipei.

Dutoit Dominique, Nugues Pierre, de Torcy Patrick. 2004. The Integral Dictionary: An Ontological Resource for the Semantic Web. *Language Resource and Evaluation Conference* (LREC). Lisbon.

Fellbaum Christiane (ed.). 1998. *WordNet: An Electronic Lexical Database*. Cambridge: The MIT press.

Grefenstette Gregory. 1998. Evaluating the adequacy of a multilingual transfer dictionary for the Cross Language Information Retrieval. *Language Resource and Evaluation Conference* (LREC). Granada.

Grishman Ralph, Hirschman Lynette, Ngo Thanh Nhan. 1986. Discovery Procedures for Sublanguage Selectional Patterns: Initial Experiments. *Computational Linguistics*. n° 12(3). pp. 205-215.

Harris Zellig. 1968. *Mathematical Structures of Language*. New York: Wiley.

Harris Zellig. 1991. *A theory of language and information: a mathematical approach*. Oxford: Oxford University Press.

Jiang Jay and Conrath David. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. *International Conference on Research in Computational Linguistics*. Taiwan.

Jones Rosie, McCallum Andrew, Nigam Kamal and Riloff Ellen. 1999. Bootstrapping for Text Learning Tasks. *IJCAI'99 Workshop on Text Mining: Foundations, Techniques and Applications*, Stockholm. pp. 52—63.

Lin Dekang. 1998. An information-theoretic definition of similarity. *International Conference on Machine Learning* (ICML). Madison.

Lin Dekang and Pantel Patrick. 2002. Concept Discovery from Text. *Computational Linguistics* (COLING). Taipei. pp. 577–583.

Lin Jimmy and Katz Boris. 2003. Q/A techniques for WWW. *Tutorial — 10th Meeting of the European Association for Computational Linguistics* (EACL). Budapest.

Mel'cuk, Igor. 1996. Lexical Functions: A Tool for the Description of Lexical Relations in the Lexicon. In L. Wanner (ed.): *Lexical Functions in Lexicography and Natural Language Processing*, Amsterdam/Philadelphia: Benjamins. pp. 37–102.

Morin Emmanuel and Jacquemin Christian. 1999. Projecting corpus-based semantic links on a thesaurus. *Association for Computational Linguistics* (ACL). Maryland. pp. 389–396.

MUC-6 (1995) *Proceedings Sixth Message Understanding Conference (DARPA)*. San Francisco: Morgan Kaufmann Publishers.

Muslea Ian. 1999. Extraction patterns for Information Extraction tasks: a survey. AAAI'99 (available at the following URL: http://www.isi.edu/~muslea/RISE/ML4IE/)

Pazienza Maria Teresa (ed.). 1997. *Information extraction*. Heidelberg: Springer Verlag (Lecture Notes in computer Science).

Paumier Sébastien. 2002. *Unitex user manual*. Manuscript. (`http://www-igm.univ-mlv.fr/~unitex/`)

Poibeau Thierry. 2003. *Extraction automatique d'information : du texte brut au web sémantique*. Paris : Hermès.

Poibeau Thierry. 2003b. The Multilingual Named Entity Recognition Framework. *European Association for Computational Linguistics Conference* (EACL). Budapest. pp. 155–158.

Ravichandran Deepak and Hovy Eduard. 2002. Learning Surface Text Patterns for a Question Answering System. *Association for Computational Linguistics* (ACL). Phildelphia. pp. 41–47

Riloff Ellen. 1995 Little Words Can Make a Big Difference for Text Classification. *Special interest Group in Information retrieval* (SIGIR). Seattle. pp. 130—136.

Riloff Ellen and Jones Rosie. 1999. Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. *National Conference on Artificial Intelligence* (AAAI). Orlando. pp. 474–479.

Silberztein Max. 1993. *Dictionnaires électroniques et analyse automatique des textes.* Paris : Masson.

Silberztein Max. 1999. Traitement des expressions figées avec INTEX. *Linguisticae Investigationes (n° spécial « Analyse lexicale et syntaxique : le système Intex »).* pp. 425—449.

Szpektor Idan, Tanev Hristo, Dagan Ido and Coppola Bonaventura 2004. Scaling web-based acquisition of entailment relations, *Empirical Methods in Natural Language Processing* (EMNLP). Barcelona.

TREC. 2000. The Ninth Text REtrieval Conference (TREC 9). Gaithersburg, 2000. (`http://trec.nist.gov/pubs/trec9/t9_proceedings.html`).

Van Rijsbergen, Cornelius J. 1979 *Information Retrieval*. London: Butterworths.

Voorhees Ellen. 1999. Natural language processing and information retrieval. In *Information extraction, toward scalable, adaptable systems.* Heidelberg: Springer Verlag (Lecture Notes in computer Science). pp. 32–48.

Wilks Yorick and Catizone Roberta. 2002. What is lexical tuning. *Journal of Semantics*. n°19(2). pp. 167-190

Yangarber Roman. 2000. *Scenario Customization for Information Extraction*. PhD Thesis, New York University.

Summary - Automatic extraction of paraphrastic phrases from small size corpora

This paper presents a versatile system intended to acquire paraphrastic phrases from a small-size representative corpus. In order to decrease the time spent on the elaboration of resources for NLP system (for example for Information Extraction), we suggest to use a knowledge acquisition module that helps extracting new information despite linguistic variation. This knowledge is semi-automatically derived from the text collection, in interaction with a large semantic network.

Keywords: Information extraction, knowledge acquisition, lexical information, semantic network, corpus.

Résumé – Extraction automatique de paraphrases à partir de petits corpus

Cet article présente un système permettant d'acquérir de manière semi-automatique des paraphrases à partir de corpus représentatifs de petite taille. Afin de réduire le temps passé à l'élaboration de ressources pour des systèmes de traitement des langues (notamment l'extraction d'information), nous décrivons un module qui vise à extraire ces connaissances en prenant en compte la variation linguistique. Les connaissances sont directement extraites des textes à l'aide d'un réseau sémantique de grande taille.

Mots clés : extraction d'information, acquisition de connaissances, sémantique lexicales, réseau sémantique, corpus.

Authors' address:

Thierry Poibeau
Laboratoire d'Informatique de Paris-Nord
Université Paris 13 and CNRS UMR 7030
99, av. J.-B. Clément
93430 Villetaneuse
Thierry.poibeau@lipn.univ-paris13.fr

Dominique Dutoit
Memodata et CRISCO
17 rue Dumont d'Urville
14000 CAEN
d.dutoit@memodata.com