

# Learning Paraphrases to Improve a Question-Answering System

**Florence Duclaye**

France Télécom R&D and ENST

2 avenue Marzin

22307 Lannion Cedex, France

florence.duclaye@rd.francetelecom.com

**François Yvon**

ENST

46 rue Barrault

75634 Paris Cedex 13, France

yvon@enst.fr

**Olivier Collin**

France Télécom R&D

2 avenue Marzin

22307 Lannion Cedex, France

olivier.collin@rd.francetelecom.com

## Abstract

In this paper, we present a nearly unsupervised learning methodology for automatically extracting paraphrases from the Web. Starting with one single linguistic expression of a semantic relationship, our learning algorithm repeatedly samples the Web, in order to build a corpus of potential new examples of the same relationship. Sampling steps alternate with validation steps, during which implausible paraphrases are filtered out using an EM-based unsupervised clustering procedure. This learning machinery is built on top of an existing question-answering (QA) system and the learnt paraphrases will eventually be used to improve its recall. We focus here on the learning aspect of this system and report preliminary results.

## 1 Introduction

Question-answering systems (Voorhees, 1999) require efficient and sophisticated NLP tools, crucially capable of dealing with the linguistic variability of questions and answers, which reflects the widely acknowledged fact that the same meaning can be conveyed using a wide variety of lexico-syntactic structures (forms). This situation is by no means specific to the QA domain, this variability being a source of difficulties for most practical applications of NLP.

Part of this variability can be captured at the syntactic level, where it takes the form of regular alternations between for instance active and passive forms, or verbal and nominal expressions of a concept. A more systematic treatment however requires some form of semantic knowledge, such

as the one found in semantic networks (Miller et al., 1990). The help provided by these resources is limited as (i) synonymy relationships found in such dictionaries cannot be taken at face value, for the lack of contextual information; (ii) synonymy implies a notion of paraphrasing which is far too restricted for our application: it is often the case that the answer to a question is expressed using terms which are only loosely (eg. metaphorically) related to the ones used in the question. For instance, "X caused Y" can be considered to be semantically similar to "Y is blamed for X" in the context of question-answering (Lin and Pantel, 2001). Rather than trying to manually complete these static resources, a virtually endless process, we have chosen to explore the benefits of a corpus-based approach and to learn such equivalences automatically. We will refer to these relationships as paraphrases, although we adopt here a rather restricted definition of paraphrase, focusing mostly on two types of linguistic phenomena: linguistic paraphrases and semantic derivations. (Fuchs, 1982) describes paraphrases as sentences whose denotative linguistic meaning is equivalent. Semantic derivations are sentences whose meaning is preserved, but whose lexico-syntactic structure is different (e.g. *AOL bought Netscape / the acquisition of Netscape by AOL*). The corpus we use for acquiring paraphrases is the Web. Using the Web as a corpus offers several clear advantages (see also (Grefenstette, 1994)): (i) it contains a great variety and redundancy: the same information is likely to occur under many guises, a property on which our learning algorithm heav-

ily relies; (ii) contextual information is available and can be used to restrict the scope of a paraphrase relationship. Moreover, as our QA system uses the Web as its only information source, it is important to extract those formulations of a given concept which are actually frequently used on the Web. This strategy is not without its own difficulties: in particular, reducing the level of noise in the acquired data becomes a serious issue. The learning mechanism we propose is capable of automatically acquiring multiple formulations of a given semantic relationship from *one single example*. This seed data consists of one instance of the target semantic relationship, where both the linguistic expression of the relationship (formulation) and the tuple of arguments have been identified. This kind of data is directly provided by our QA system, but is also widely available in usual dictionaries. Given this positive example, our learning machinery repeatedly queries the Web, trying alternately to use the currently known formulations to acquire new argument tuples, and the known argument tuples to find new formulations. This mechanism decomposes into two steps: the search for potential paraphrases of the semantic relation and the validation of these paraphrases, which is based on frequency counts and the Expectation-Maximisation (EM) algorithm.

This paper introduces, in Section 2, some background technical work which has been influential for our approach, as well as related research on paraphrase learning. Section 3 then gives a thorough presentation of our system, first giving a general overview of its behavior, then explaining our EM-based filtering strategy, and finally going into the details of the acquisition procedure. Before concluding, we discuss in Section 4 some experimental results that highlight the interest of our approach.

## 2 Background

### 2.1 Paraphrase learning

As paraphrases can be used in various contexts and applications, learning them is accomplished using very different methodologies. (Barzilay and McKeown, 2001) distinguish between three different methods for collecting paraphrases. The first

one is manual collection, the second one is the use of existing linguistic resources, and the third one is corpus-based extraction of similar words or expressions. Of these three methods, manually collecting paraphrases is certainly the easiest one to implement, though probably the most tedious and time-consuming one.

Linguistic resources such as dictionaries can prove to be useful for collecting or generating paraphrases. For instance, (Kurohashi and Sakai, 1999) uses a manually-tailored dictionary to rephrase as verbal phrases ambiguous noun phrases. Such linguistic resources as dictionaries may be useful for disambiguation purposes, but they rarely provide linguistic information in context, so that the proper paraphrases cannot always be spotted. Moreover, they are often recognised to be poorly adapted to automatic processing (Habert et al., 1997). (Torisawa, 2001) proposes a method using the Expectation-Maximisation algorithm to select verb schemes that serve to paraphrase expressions.

Finally, some of the works in the area of corpus-based extraction of similar words or expressions rely on Harris' Distributional Hypothesis, stating that words occurring in the same context tend to have similar meanings. Relying on this assumption, (Barzilay and McKeown, 2001) and (Akira and Takenobu, 2002) work on a set of aligned texts and use contextual cues based on lexical similarities to extract paraphrases. In the same line, (Lin and Pantel, 2001) uses an unsupervised algorithm for discovering inference rules from text. Instead of applying Harris' rule to words, the authors apply it to paths in dependency trees of a parsed corpus.

### 2.2 Information extraction by bootstrapping

Recent work on information extraction provides us with interesting approaches that can be adapted to solving the problem of paraphrase learning. (Riloff and Jones, 1999) describes an information extraction system relying on a two-level bootstrapping mechanism. The "mutual bootstrapping" level alternatively constructs a lexicon and contextual extraction patterns. The "meta-bootstrapping" level keeps only the five best new terms extracted during a given learning round before continuing with the mutual bootstrapping. In this way, the

author manages to reduce the amount of invalid terms retrieved by the application of extraction patterns.

The DIPRE technique (Dual Iterative Pattern Relation Extraction) presented in (Brin, 1998) is also a bootstrapping method, used for the acquisition of (author,title) pairs out of a corpus of Web documents. Starting from an initial seed set of examples, the author constructs extraction patterns that are used to collect (author,title) pairs. In their turn, these pairs are searched in the corpus and are used to construct new extraction patterns, and so on. Finally, (Collins and Singer, 1999) describes a method for recognising named entities with very little supervision data by building two classifiers operating on disjoint feature sets in parallel.

### 3 System overview

#### 3.1 General overview of the paraphrase learning system

Our paraphrase inference algorithm learns from one single positive example, using a two-level bootstrapping mechanism. This seed example is an answer to a question, returned by our QA system. In our model, a meaning is represented as the association between the linguistic formulation  $f$  of a predicate, and its arguments tuple  $a$ . For instance, one example of the “authorship” relationship would be represented as:  $f=$ ”to be the author of”,  $a=$ (”Melville”, ”Moby Dick”). Identification of paraphrases relies on a probabilistic decision model, whose parameters are estimated in an almost unsupervised way. Estimation relies on an EM-based clustering algorithm presented in Section 3.2: it takes as input a matrix containing frequency data for the co-occurrence of a set of formulations  $F$  and the corresponding argument tuples  $A$ , as measured in a corpus  $C$ .

Our initial corpus  $C_i$  contains one unique “seed” example expressing the target relationship, and represented as the cooccurrence of a formulation  $f_i$  and an argument tuple  $a_i$ . Given this seed, we would like to build a new corpus  $C$ , potentially containing many more instances of the target relationship. This is done by using independently  $f_i$  and  $a_i$  to formulate queries, which are used to sample from the web. The retrieved documents

are searched for new interesting formulations and arguments pairs, repeatedly used to produce new queries, which in turn will extract more arguments and formulations... During this stage, we need to be able to (i) generate queries and process the retrieved documents so as to (ii) extract new formulations and argument tuples. Details of these corpus building procedures are given in Section 3.3.

The quality of the extracted paraphrases depends critically on our ability to keep the expanding corpus *focused on the target semantic relationship*: to this end, the acquisition phases are interleaved with filtering stages, which are also based on our EM-based clustering. Filtering is indeed critical to ensure the convergence of this procedure. The overall architecture of our system is represented on figure 1.

#### 3.2 Filtering with the Expectation-Maximisation algorithm

The filtering problem consists in sorting out incorrect paraphrases of the original relationship from valid ones. This amounts to classifying each formulation in our corpus as 1 (valid paraphrase) or 0 (not valid), based on co-occurrence data between arguments tuples and formulations. This bipartitioning problem is weakly supervised, as we initially have one positive example: the seed formulation. This is a favorable configuration for the use of EM-based clustering algorithms for co-occurrence data (Hofmann and Puzicha, 1998). We thus assume that each phrase (consisting of a formulation  $f$  and its arguments  $a$ ) is generated by the following stochastic model:

$$P(f, a) = \sum_{s \in S} P(f, a | s) P(s) \quad (1)$$

$$= \sum_{s \in S} P(f | s) P(a | s) P(s) \quad (2)$$

where  $S$  is the set of semantic relationships expressed by sentences in our corpus. We further assume that our corpus only contains two such relationships, whose values are defined as  $S = 1$ , meaning that a given sentence expresses the same relationship as the seed sentence, and  $S = 0$ , meaning that the sentence expresses another (unspecified) relationship.

Given this model, the reestimation formulas are easily derived (see eg. (Hofmann and Puzicha,

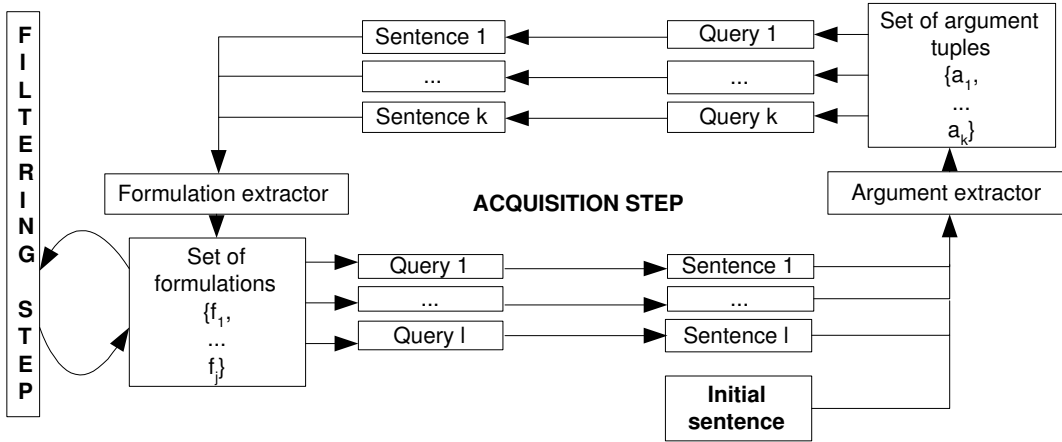


Figure 1: Paraphrase learning system

1998)); they are given in Table 1, where  $N()$  denotes the count function.

### E-Step

$$P(s|f, a) = \frac{P(s)P(f|s)P(a|s)}{\sum_i P(s_i)P(f|s_i)P(a|s_i)} \quad (3)$$

### M-Step

$$P(a|s) = \frac{\sum_{f \in F} N(f, a)P(s|f, a)}{\sum_{a \in A} \sum_{f \in F} N(f, a)P(s|f, a)} \quad (4)$$

$$P(f|s) = \frac{\sum_{a \in A} N(f, a)P(s|f, a)}{\sum_{f \in F} \sum_{a \in A} N(f, a)P(s|f, a)} \quad (5)$$

$$P(s) = \frac{\sum_{f \in F} \sum_{a \in A} N(f, a)P(s|f, a)}{\sum_{f \in F} \sum_{a \in A} N(f, a)} \quad (6)$$

Table 1: Reestimation formulas for EM

This model enables us to incorporate supervision data during the parameter initialisation stage, where we use the following values:  $P(S = 1|f_i, a_i) = 1$  and  $P(S = 1|f_i, a) = 0.6, \forall a \neq a_i$  in equation (3). All the other values of  $P(S|F, A)$  are taken equal to 0.5. EM is then run until convergence of the maximised parameters. In our case, this convergence generally achieved within 10 iterations.

Once the parameters have been learnt, we use this model to decide whether a formulation  $f$  is a valid paraphrase based on the ratio between  $P(S = 1|f)$  and  $P(S = 0|f)$ , computed as:

$r = \frac{P(S=1)P(f|S=1)}{P(S=0)P(f|S=0)}$ . Given that  $P(S = 1)$  is grossly overestimated in our corpus, we require this ratio to be greater than a predefined threshold  $\theta > 1$ .

### 3.3 The acquisition procedure

The main tool used during the acquisition step is our QA system itself, which has been adapted in order to be also used as an information extraction tool. The original system has two main components. The first one turns an input question into a Web query and performs the search. The second module analyses the retrieved pages, trying to match answers, an answer corresponding to a set of predefined extraction patterns. Both the query and the extraction patterns are derived from the original question using rules. Details regarding this QA system and the NLP components involved at each stage are given in (Duclaye et al., 2002).

In “learning” mode, we by-pass the query construction phase and enforce the use of the argument tuples (or formulations) as search keywords. The analysis phase uses very general information extraction patterns directly derived from the arguments (or formulations) being processed. Assume, for instance, that we are in the process of searching for paraphrases, based on the argument pair [“Melville”, “Moby Dick”]. Both arguments will be used as keywords, and two patterns will be matched in the retrieved documents : “Melville [Verb] Moby Dick” and “Moby Dick

[Verb] Melville”. In this example, a verb is required to occur between the two keywords. This verb will be considered to be a potential paraphrase of the initial formulation. For each query, only the top N documents returned by the search engine are considered.

Notwithstanding the effects of the filtering procedure, the extracted (*arguments, formulations*) are cumulated, round after round, in a corpus  $C$ , from which statistics are then estimated. This iterative process of acquiring formulations and argument tuples, combined with the validation process at the end of every iteration, converges and ends up when no new formulation is found.

#### 4 Experimental results

The experiments described in this section were conducted on 18 initial sentences, representing 12 different semantic relationships (e.g. purchase of a company, author of a book, invention of something, ...). See table 2 for examples of formulations and argument tuples. For each of these sentences, the learning procedure described in Section 3 was run on one iteration. The results presented here were obtained by searching for French documents on the Web and taking the first N=1000 previews returned by the search engine.

The extracted paraphrases were manually checked and classified as valid or invalid by ourselves. In this application, success is only measured as the average precision of the extracted paraphrases which should eventually be fed into the QA system. Recall, in comparison, is unimportant, as long as we can find the most frequently used paraphrases. The selection ratio represents the percentage of formulations classified as valid paraphrases by our system. The decision to classify a formulation as a valid or invalid paraphrase is based on the ratio between  $\log(P(S = 1|f))$  and  $\log(P(S = 0|f))$ , called  $\theta$ . The selection ratios and precision results for various filtering thresholds  $\theta$  are reported in table 3.

In these experiments, the best average precision achieved is 66.6%, when  $\theta = 186$ . Performed on several relationships, these experiments showed that the precision rate may vary importantly from one semantic relationship to another : it can be

| theta | selection ratio | precision |
|-------|-----------------|-----------|
| 7     | 44.0%           | 42.9%     |
| 25    | 29.8%           | 47.3%     |
| 48    | 23.9%           | 47.3%     |
| 117   | 14.2%           | 54.9%     |
| 186   | 10%             | 66.6%     |
| 232   | 9.4%            | 65.4%     |

Table 3: Experimental results

as high as 100% for certain relationships, and as low as 6% for others. These results may seem to be low. This is partly due to the varying amount of data extracted from the Web for the semantic relationships. Applying the same threshold  $\theta$  to all relationships may not be the best method, as the system extracts a huge quantity of formulations for certain relations, and a small one for others. Moreover, the majority of the formulations wrongly classified as good paraphrases are thematically related to the seed formulation (e.g. for the purchase relationship : to own, to belong, to merge, ...).

As indicated in table 3, the increasing values of  $\theta$  cause the selection ratios to decrease and the precision to increase. The general tendency is that as  $\theta$  gets bigger and bigger, the amount of formulations classified as bad paraphrases increases, so that eventually only the seed formulation is kept as valid. Increasing  $\theta$  is thus insufficient to improve the average precision of the extracted paraphrases. A balance needs to be found between the selection ratio and the precision of the extracted paraphrases.

Let us point out that the results shown in table 3 only reflect the first learning iteration. Other experiments were conducted on several learning iterations, which lead us to believe that precision should increase with the number of iterations. Table 4 shows the results obtained on the purchase relationship, after five learning iterations. The filtering strategy was different from the one detailed in Section 3.2. Instead of keeping the formulations according to the ratio between  $P(S = 1|f)$  and  $P(S = 0|f)$ , we decided to only keep the five best formulations at each learning iteration.

|                  |                          |   |
|------------------|--------------------------|---|
| purchase of      | ”acheter” (to buy)       | AOL; Netscape                                     |
| author of        | ”écrire” (to write)      | Melville; Moby Dick                               |
| inventor of      | ”inventer” (to invent)   | Gutenberg; imprimerie ( <i>printing machine</i> ) |
| assassination of | ”assassiner” (to murder) | Oswald; Kennedy                                   |

Table 2: Some exemplary relationships and formulations

| Iter. | Formulations classified as valid paraphrases  |
|-------|---|
| 1     | racheter ( to buy out), acquirir (to acquire), acheter (to buy), utiliser (to use), recevoir (to receive) |
| 2     | racheter, acquirir, acheter, reprendre (to take back), absorber (to take over)                            |
| 3     | racheter, acheter, acquirir, qui racheter ([which] buy out), devenir (to become)                          |
| 4     | racheter, acheter, acquirir, absorber, grouper (to gather)  |
| 5     | racheter, acheter, reprendre, devenir, acquirir   |

Table 4: Results of five learning iterations on the purchase relationship

## 5 Conclusions and future work

In this paper, we have presented a nearly unsupervised methodology for learning paraphrases automatically, starting with one single positive learning example. Using an EM-based validation strategy, we are able to filter out the invalid potential paraphrases extracted during the acquisition steps.

Not only are these paraphrases useful to improve the results of our question answering system, but the acquired argument tuples could also be used for other purposes than paraphrase learning, such as the construction of semantic lexicons. In fact, the filtering step could as well be applied on the acquired argument tuples.

Beyond its promising experimental results, the adaptability of our approach brings to the fore the multiple practical applications of this work. Focused on the acquisition and validation steps, various improvements are presently under investigation. Concerning the acquisition step, we are planning to learn multilingual paraphrases, as well as more complex extraction patterns (involving nominalisations). We are also considering using automatically learnt contextual information to refine the quality of the queries we use to sample the Web. Future improvements of the filtering / validation step will aim at testing other filtering strategies.

Based on a language-independent learning strategy, our paraphrase learning system will be integrated into the multilingual question-answering system. Our system will act as an offline com-

ponent which will learn paraphrases of answers returned by the QA system. Its integration will not require many developments, as the QA system already takes into account manually-entered paraphrasing rules. We will thus have to automate this process of entering paraphrasing rules into the QA system. This integration will enable us to evaluate our methodology and to measure the improvements incurred by this paraphrase learning module.

## References

- Terada Akira and Tokunaga Takenobu. 2002. Automatic disabbreviation by using context information. In *Proceedings of the NLPRS Workshop on Automatic Paraphrasing : Theories and Applications*.
- Regina Barzilay and Kathleen R. McKeown. 2001. Extracting paraphrases from a parallel corpus. In *Proceeding of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 50–57, Toulouse.
- Sergei Brin. 1998. Extracting patterns and relations from the world wide web. In *Proceedings of WebDB Workshop at EDBT*.
- Michael Collins and Yoram Singer. 1999. Unsupervised models for named entity classification. In *Proceedings of the Workshop on Empirical Methods for Natural Language Processing*.
- Florence Duclaye, Pascal Filoche, Jerzy Sitko, and Olivier Collin. 2002. A polish question-answering system for business information. In *Proceedings of the Business Information Systems Conference, Poznan*.
- Catherine Fuchs. 1982. *La Paraphrase*. Presses Universitaires de France.

- Gregory Grefenstette. 1994. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, Boston.
- Benoît Habert, Adeline Nazarenko, and André Salem. 1997. *Les linguistiques de corpus*. Armand Colin, Paris.
- Thomas Hofmann and Jan Puzicha. 1998. Statistical models for co-occurrence data. Technical Report AI. 1625, MIT, AI Lab.
- Sadao Kurohashi and Yasuyuki Sakai. 1999. Semantic analysis of japanese noun phrases : a new approach to dictionary-based understanding. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 481–488.
- Dekang Lin and Patrick Pantel. 2001. Discovery of inference rules for question-answering. In *Natural Language Engineering*, volume 7, pages 343–360.
- George Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine Miller. 1990. Introduction to wordnet: An on-line lexical database. In *Journal of Lexicography*, volume 3, pages 234–244.
- Ellen Riloff and Rosie Jones. 1999. Learning dictionaries for information extraction by multi-level bootstrapping. In *Proceedings of the 16th National Conference on Artificial Intelligence*.
- Kentaro Torisawa. 2001. A nearly unsupervised learning method for automatic paraphrasing of japanese noun phrases. In *Proceedings of the NLPRS 2002 workshop on Automatic Paraphrasing : Theories and Applications*, Tokyo.
- Ellen Voorhees. 1999. The TREC-8 question answering track report. In *Proceedings of TREC-8*.