Using the Web as a Linguistic Resource for Learning Reformulations Automatically

Florence Duclaye*†, François Yvon†, Olivier Collin*

* France Télécom R&D, 2 avenue Pierre Marzin, 22307 Lannion Cedex, France
{florence.duclaye, olivier.collin}@rd.francetelecom.com

†ENST (Ecole Nationale Supérieure des Télécommunications)
46 rue Barrault, 75634 Paris Cedex 13, France
{yvon,duclaye}@enst.fr

Abstract

The use of paraphrases as a potential way to improve question answering, machine translation or automatic text summarization systems has long attracted the interest of researchers in natural language processing. However, manually entering reformulations into a system is a tedious and time-consuming process, if not an endless one. In this paper, we introduce a learning machinery aimed at acquiring reformulations automatically. Our system uses the Web as a linguistic resource and takes advantage of the results of an existing question answering system. Starting with one single prototypical argument tuple of a given semantic relation, our system first searches for potential alternative formulations of the relation, then finds new potential argument tuples, and iterates this process to progressively validate the candidate formulations. This learning process combines an acquisition stage, whose goal is to retrieve new evidences from Web pages, and a validation stage, whose role is to filter out noise and discard invalid paraphrases. After justifying the use of the Web as a linguistic resource, we describe our system, and report on primary results on a series of test semantic relations.

1. Introduction

The study of paraphrases, as a potential means to improve question answering, machine translation or automatic text summarization systems has long attracted the interest of researchers in natural language processing. Considering the number of current works devoted to reformulation using, one realises that reformulations are real stakes: (Mitamura and Nyberg, 2001) use reformulations for automatic rewriting in the field of controlled language translation; (Ohtake and Kazuhide, 2001) explore the impact of paraphrasing honorifics; (Tomuro and Lytinen, 2001) proposes a method to select features and paraphrase questions; (Dras, 1998) describes a model where paraphrases are used to enforce textual surface constraints (such as length, …), based on decision rules and change measurements. Let us point out that the focus here is on the use of paraphrases. That is the reason why none of the works on translation is quoted in this section. Despite the vast amount of works dealing with paraphrase using, entering reformulations or reformulation rules manually into a system is a tedious and time-consuming task, if not an endless one. Our goal is thus to automate the process of acquiring reformulations from corpora, so as to improve a general purpose question answering (QA) system (Duclaye et al., 2002).

Given the context of our application, we have adopted a rather limited definition of reformulations, and have concentrated solely 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. Melville wrote Moby Dick / Melville is Moby Dick's author). For similar reasons, we have favoured the use of the world wide web, which is indeed the knowledge source used by our QA system, and where we expect to find the most useful reformulations for contributing to improving the precision and recall of the QA system.

The learning mechanism we have implemented works in an unsupervised fashion, and is able to automatically acquire multiple formulations of a given semantic relation from one single example. The seed data consists of one instance of the target relationship, where both the linguistic expression of the relationship and the tuple of arguments have been properly identified. This kind of data is directly provided by our question answering system. Given this unique positive example, our learning machinery repeatedly queries the Web, trying alternatively to use the currently known formulations to acquire new potential arguments, and the known arguments to find new formulations. This mechanism decomposes into two stages: the search for potential reformulations of the semantic relation under study and the validation or invalidation of these potential reformulations, which aims at filtering out the noise collected during the search stage, based on frequency counts.

This paper is organised as follows. Section 2 describes the current background in reformulation learning. This section also presents recent works in information extraction that can be considered to be the founding statements of our approach to reformulation learning. Section 3 discusses the use of the web as a linguistic resource. Section 4 provides an overview of our question-answering system, mainly emphasizing the infrastructure components that are useful for extracting reformulations, and describes in some detail our learning methodology. Before suggesting directions for future work and providing concluding remarks (section 6), section 5 reports on some preliminary experimental results that highlight both the interest of this approach, and underline its practical difficulties.

2. Background

2.1. Reformulation learning

(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. The third one is corpus-based extraction of similar words or expressions. Of these three methods, manual collection of paraphrases is certainly the easiest one to implement, though probably the most tedious one. Linguistic resources such as dictionaries or semantic networks can also prove useful for collecting or generating paraphrases. Along these lines, (Boyer and Lapalme, 1985) describe a method for generating paraphrases from meaning-text semantic networks and lexical rules. (Kurohashi, 1999) uses a manually tailored dictionary to rephrase as verbal phrases ambiguous noun phrases of the form "word1 no word2" ("no" is a Japanese postposition which can express a great variety of semantic relations). Furthermore, in the past few years, more and more works have been focusing on paraphrase collecting via corpus-based extraction. As a corpus, (Barzilay and McKeown, 2001) use a set of aligned texts, all of which are translations of one document. Their founding idea is that words appearing in similar contexts in aligned sentences are very likely to be paraphrases. Based on this assumption, these authors use contextual cues based on lexical similarities to extract paraphrases. (Sekine, 2001) presents a very similar methodology: the corpus the author uses is a set of newspaper articles published on the same day. (Akira and Takenobu, 2001) describe yet another example of approach to paraphrase collecting via corpus-based extraction. The authors' goal is to find semantic equivalences of abbreviations and acronyms. To achieve this purpose, a collection of texts dealing with the aviation domain was chosen, which contains 11% of abbreviations and 2% of acronyms. Similarly to (Barzilay and McKeown, 2001), this work assumes that words appearing on each side of an abbreviation or of an acronym are statistically similar to those appearing on each side of the corresponding full form. Finally, (Torisawa, 2001) proposes to use the Expectation-Maximisation (EM) algorithm to select verb schemes that serve to paraphrase expressions of the form "word1 no word2" (cf Kurohashi's work description, above).

2.2. Information extraction

Recent work on information extraction provides us with interesting approaches that can be adapted to solving the problem of reformulation learning. (Riloff, 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, S. Brin 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) uses a corpus of some 970,000 sentences from the New York Times, for the purpose of learning to recognise named entities in a nearly unsupervised fashion. Each training instance is divided into two parts, deemed to be equally informative for the task at hand: a (feature-based representation of) the spelling of the named entity, and a (feature-based representation of) its local syntactic context. This dual representation allows the authors to use a bootstrapping method, building two classifiers in parallel. Starting with a handful of positive examples of named entities of three different kinds (person, organisation or location), the system enters a series of induction rounds, alternatively using either the contextual or the spelling features for describing the training instances. After each round, the learnt classifier is used to label more data, which is added as supplementary training data for the next learning round.

3. The Web as a linguistic resource

This part aims at explaining why we chose to learn reformulations from documents found on the Web, rather than resorting to closed corpora, or to lexical resources such as dictionaries or thesauri. In (Habert et al., 1997), John Sinclair defines a corpus as a "collection of language data which are selected and organized according to explicit linguistic criteria, in order to be used as a language sample". This part focuses on the linguistic criteria needed in our corpus to learn reformulations. Despite some drawbacks (discussed at the end of the present part), the Web is considered to be the most adequate corpus when considering these criteria for reformulation learning.

3.1. Improving question answering on the Web

The first criterion that the corpus needs to satisfy relates to the very purpose of reformulation learning: improving our question answering system. We need to find reformulations that will help our QA system to find answers in Web documents: it thus seems appropriate to look for formulations that actually occur in these kinds of documents. Additionally, the Web provides the best approximation to date of a constantly updated source of linguistic knowledge, instantaneously assimilating changes in language use. This point is already made in (Habert et al., 1997) which states that the vocabulary of network navigation is permanently changing. As an example, the newly-born French word "globalisation" (globalisation), which is a synonym of French "mondialisation" does not appear in the electronic synonym dictionary of the University of Caen. Again it is felt that our QA system can only benefit from integrating these changes, as they might well be used by users of the QA system, which is itself a service accessible on the Internet.

3.2. Variety and redundancy

In the context of our learning methodology, the Web presents two additional advantages: variety and redundancy. Variety is a concern, since we want to be able to acquire new formulations for virtually any semantic relationship: this fact alone precludes the use of specialised corpora. In fact, not only can we find information on virtually any subject on the Web, but it often seems to be the case that the same piece of information appears in multiple Web documents, under multiple linguistic guise and in several languages. Such inherent redundancy leads to a fantastic linguistic variety; both qualities being in fact expected by our learning methodology, as will appear clearly in section 5. To get a hint at this amazing variety, just consider the two following verbalisations found on the Web which both express the semantic relationship "company 1 [buy] company 2": "l'acquisition de Netscape par AOL" (the acquisition of Netscape by AOL) and "IBM croque Informix Software" (IBM eats Informix Software). While the first one is formal style, the second one is much more familiar and metaphoric and is very unlikely to be found in any kind of general purpose thesaurus.

3.3. Linguistic information in context

Lexical resources, such as dictionaries or semantic networks, describe words and concepts, focusing on their different meanings, relations and uses. Although dictionaries are useful for disambiguation purposes, they are often recognized not to be adapted to automatic processing (Habert, 1997). Thesauri, for their part, can provide information such as synonyms, antonyms or hyperonyms of terms. They seem to be very suited for providing us with interesting paraphrases. This is not always the case, for thesauri often lack the kinds of contextual information necessary to spot the proper reformulations. To illustrate this point, an experience was conducted on the synonyms of the verb "acheter" (buy) in the sentence "AOL a acheté Netscape" (AOL bought Netscape). Among the 29 synonyms proposed by the online synonym dictionary of the University of Caen (http://elsap1.unicaen.fr/cherches.html), one of them is indeed very interesting: "acquérir" (to acquire). However, replacing "acheter" with other synonyms such as "corrompre" (to corrupt) or "importer" (to import) in our test sentence does not make sense. The point is that even if "corrompre" and "importer" are indeed possible synonyms of the verb "acheter", they are only acceptable synonyms in restricted contexts. For instance, "l'Angleterre achète du café du Brésil" (England buys coffee from Brazil) may be paraphrased into "l'Angleterre importe du café du Brésil" (England imports coffee from Brazil), but not into "l'Angleterre acquiert du café du Brésil" (England acquires coffee from Brazil). In (Pirrelli and Yvon, 1999), F. Yvon and V. Pirrelli indeed explain that "learning the correspondence between any two linguistic levels involves classifying one unit *a* of A as one unit *b* of B *given a certain context*". Consequently, replacing one word with its synonyms provided by a thesaurus does not always yields acceptable reformulations. In contrast, the paraphrases we find on the Web always come with a context, which may help to reduce the risk of over-generalising the scope of synonymy relationships

3.4. Noise

Digging information from the Web however implies a number of practical difficulties, most of which are direct consequences of its heterogeneity. Web documents are both heterogeneous in nature (technical reports, newspaper articles, advertisements, literary documents, etc), in content, in style and in editing quality (most documents are not error-prone, both from a syntactic and a lexical point of view). The consequences for our learning methodology are twofold: (i) all the information found on the Web cannot be taken at face value, which implies that information sources need to be cross-validated; (ii) linguistic formulations also need to be somewhat double-checked so as to avoid acquiring invalid patterns. Due to the Web's heterogeneity, reducing the level of noise becomes a serious issue, which we address both by using a combination of robust natural language processing tools and statistical routines.

To conclude, the type of corpus needed for reformulation learning is a vast reference corpus of written text, containing a maximal amount of variety and redundancy, both from the point of view of form and content. Lexical corpora certainly are interesting linguistic information resources, but their lack of contextual information makes them less suited to our needs. Given finally that our aim is to improve an online QA system, the Web appears, despite its drawbacks, to be the place where one should look for inducing useful reformulations.

4. System overview

4.1. Question-answering system

WebStorm is a real time multilingual question answering system that relies on the Web to find the answers (Duclaye et al., 2002). It is designed to answer factual natural language questions. The system works as follows (see figure 1). In a first stage, the input question is analysed by a general purpose linguistic analysis system, which assigns morpho-syntactic tags and identifies coherent syntactic units (chunks). This analysis stage results in two kinds of outputs: (i) keywords, which are extracted from the question and further used to query traditional keyword-based search engines; (ii) abstract patterns, which correspond to possible linguistic expressions of the answer. These patterns are derived through the application of hand-crafted rules, which transform a general syntactic representation of the question into a possible pattern of answer. For instance, a typical rule (albeit a quite naive one) might be:

Who(pronoun) <V>(Verb)> <X>(Noun Phrase)?
--> <Answer>(NP) <V>(Verb) <X>(NP) OR
<X>(NP) <V>(Verb) by(Prep) <Answer>(NP)

These generic linguistic rules can be backed by more specific rules (domain specific rules), making it possible for instance to retrieve « A is the author of X » in response to «Who wrote X? ». In a second stage, the documents returned by the search engine are first preprocessed (removal of non textual sections, filtering of HTML tags, segmentation in coherent textual paragraphs), then analysed using the same grammatical chunker and finally scanned for occurrences of potential answers, using the patterns obtained in (ii). Once possible answers have been identified, the system eventually ranks them according to their distance to the syntactic pattern and their frequency. Inserted items such as modifiers (e.g. adjectives, time information) do not prevent answers from being extracted, the pattern matching process being flexible enough to allow for such insertions.

This mechanism also makes it possible to define specific data sources and a non-linguistic extraction strategy for particular questions whose answers are unlikely to be found as a sentence, but more as a figure in a table (a common example being the retrieval of stock quotes). From the user`point of view, the same system will answer to «When did France Telecom buy Orange? » (answer found in a sentence) as well as «How much are France Telecom stocks today? » (answer found in a table). Figure 1 describes the architecture of our question-answering system.

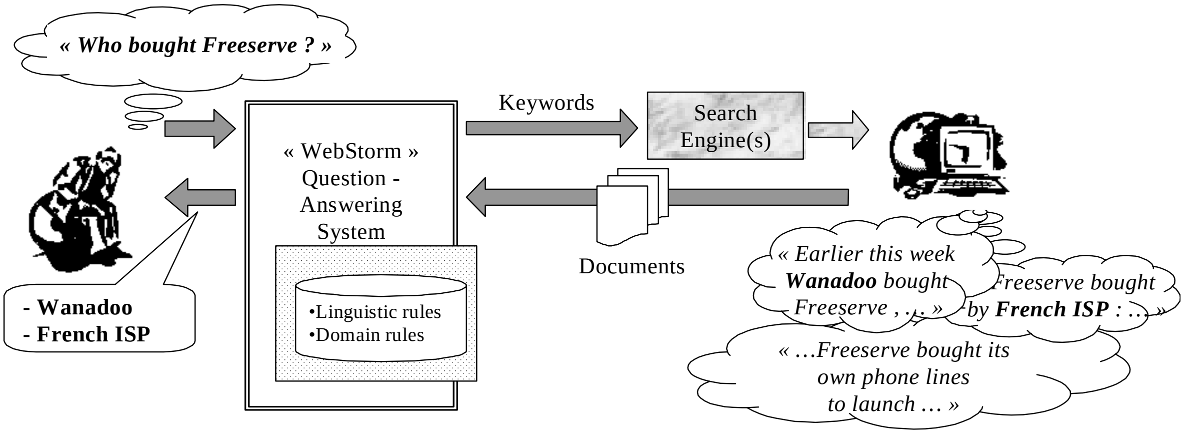
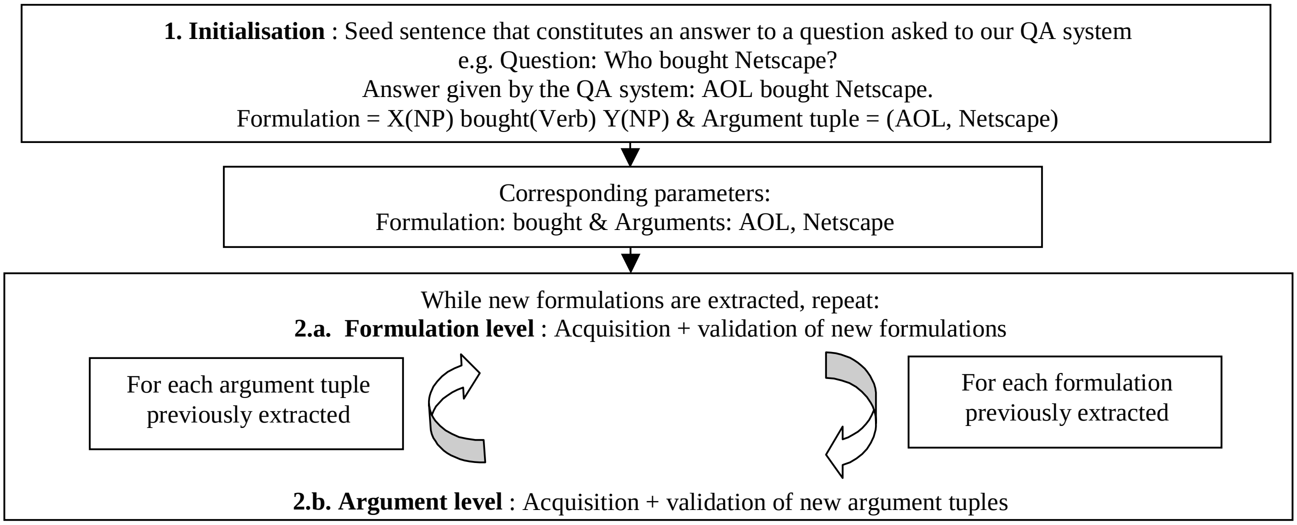

« Who bought Freeserve ? »
Keywords
Search Engine(s)
« WebStorm » Question - Answering System
•Linguistic rules
•Domain rules
Documents
- Wanadoo
- French ISP
« Earlier this week Wanadoo bought Freeserve , … »
Freeserve bought by French ISP : … »
« …Freeserve bought its own phone lines to launch … »

Figure 1: WebStorm Question-Answering system

As explained above, our system already integrates a very limited reformulation mechanism, based on hand-coded rules, which apply during the construction of extraction patterns. These resources however only exist for a handful of already seen questions. Our goal is thus to increase the number of such reformulations in an automatic fashion so as to improve the overall performance of the QA system.

4.2. Reformulation learning system

In order to learn reformulations, our system is initialised with one single positive example of a semantic relation, and then relies on a two-level bootstrapping mechanism. Figure 2 illustrates the algorithm. The reformulation learning process is composed of an acquisition and a validation step which both can be instantiated in various ways, making this methodology quite general.



1. Initialisation : Seed sentence that constitutes an answer to a question asked to our QA system
e.g. Question: Who bought Netscape?
Answer given by the QA system: AOL bought Netscape.
Formulation = X(NP) bought(Verb) Y(NP) & Argument tuple = (AOL, Netscape)

Corresponding parameters:
Formulation: bought & Arguments: AOL, Netscape

While new formulations are extracted, repeat:
2.a. Formulation level : Acquisition + validation of new formulations

For each argument tuple previously extracted

For each formulation previously extracted

2.b. Argument level : Acquisition + validation of new argument tuples

Figure 2: Reformulation learning system

The tool currently used for the acquisition step is simply the QA system itself, which has been adapted so as to become capable of digging the Web for linguistic information. In this 'learning mode', it is in fact possible to (i) by-pass entirely the keyword extraction phase and enforce the use of the arguments (or formulation) under focus as keywords; (ii) use very general information extraction pattern directly derived from the arguments (or formulation) being processed. Assume, for instance, that new formulations are being searched for based on the argument tuple [AOL, Netscape], then these arguments will be used as keywords, and two answer patterns will be searched for in the retrieved documents: `AOL [verb] Netscape` and `Netscape [Verb] AOL`. In this example, a verb is required to occur between the two keywords. This verb will define a new potential formulation of the initial semantic relation. To complete the description of this step, it is worth mentioning that, for each query, we only consider the top N documents returned by the search engine.

In essence, the validation stage aims at making a binary decision regarding the formulations or argument tuples retrieved during the previous extraction step, and discriminate between valid and invalid expressions (during step 2.a) or arguments (during step 2.b) of the original semantic relationship. Let us focus for instance on step 2.a : this step takes as input an existing set of formulations $\{f_j\}_{,\ j=1..k}$ and an existing set of argument tuples $\{t_i\}_{i=1..l}$ extracted during the previous stages. For every new candidate formulation f, cooccurrence patterns of the pairs (f, t_i) are acquired and compared with the cooccurrence patterns of the known formulations (f_j, t_i). Provided that enough evidence is found, the new formulation is added in the pool of known paraphrases, otherwise it is discarded. The decision made during this step can be quite conservative, as we are much more concerned with precision than with recall: missing a formulation is of no consequence for the next induction rounds, whereas picking up an erroneous one can easily make the algorithm diverge. It is worth adding that this model easily leans itself to a probabilistic extension: rather than making hard decision on an argument tuple or a formulation, we would just compute probability estimates.

As such, our methodology more clearly reveals its similarity to the one proposed in (Singer and Collins, 1999). Two formulations (the same holds for argument tuples) will be found similar and representative of the same class of valid formulation of the original relationship if they occur in a sufficiently large number of contexts; the main difference being that we use here a narrower definition of the context, which is in fact restricted to the argument pair. As an alternative implementation of this validation step, we are considering having the validation step take place simultaneously for a whole set of formulation and argument tuples: leaving steps 2.a and 2.b unchanged, the validation will operate on a complete cooccurrence table and simultaneously sort out the most likely paraphrases and argument pairs, using for instance the EM algorithm (Hofman and Puzicha, 1998).

5. Experimental results

In this section, we describe preliminary results of a pilot study conducted with a handful of semantic relations. Table 1 displays a series of formulations and argument tuples automatically extracted for the relation `kill(X,Y)`. These results were obtained on a relatively small amount of data as for each query (either argument tuple or formulation), we have only considered the first 100 Web documents returned by the QA system. Moreover, our extraction rules for the formulation acquisition steps were quite simple, insofar as we first chose to only extract the verbs between the arguments.

Formulations	Argument tuples
tuer (kill)	Lee Harvey Oswald – Kennedy
assassiner (assassinate)	Lincoln – John Wilkes Booth
vouloir (want)	Bill Gates – a Los Angeles
dominer (dominate)	Lee Harvey Hoswald – Kennedy
innocenter (innocenting)	Anna Kournikova – Jennifer Capriati
	Matta - papis
	Chris Carpenter – le Yankees

Table 1: Acquisition of potential reformulations for the semantic relation `kill(X,Y)` - French seed sentence: `Lee Harvey Oswald tua Kennedy`

In spite of a delibetately simplistic experimental setting, which does not fully take advantage of the redundancy of the Web, our algorithm seems to exhibit a consistent behaviour reflected in the lists in Table 1. On the 'formulation side', we can notice that we were able to learn litteral paraphrases of the original sentence. At first glance, one can notice the surprisingly low redundancy of the data in these tables. More redundant data were indeed expected to be found on such a vast corpus. The main reason for this is that we only extracted data from the first 100 Web documents returned by the question answering system. The other possible reason is that our extraction patterns may be too restrictive. We need to allow for insertions between formulations and arguments, provided these insertions do not alter the semantic relation between them. One other important remark is that we chose to start learning reformulations by extracting formulations that were strictly verbs between the arguments.

Table 1 also gives a hint at the problem of noise: it includes various formulations that cannot be considered as valid expressions of the relationship under focus (eg. “dominer”(*dominate)*), but which account for various metaphoric usage. The typography in Oswald’s name illustrates yet another case of noisy data. We need to reduce this noise by statistically and semantically filtering the data acquired.

These experiments also reveal that the same semantic relationship can in fact be expressed using either very narrow or extremely broad terms. A typical example was obtained in our experiment on questions concerning the heights of monuments: both the quite specific “culminer”(*culminate)* and the extremely ambiguous “faire”(*do*) were acquired from the seed sentence “La tour Eiffel mesure 300m”(*the Eiffel tower is 300 meters high*). If used during step 2.b, these general formulations increase the amount of retrieved documents importantly, and flood the system with inaccurate argument tuples. A first remedy might be to discard these formulations altogether (based for instance on a stop list of overly ambiguous terms). Given their frequency in texts, we feel that such formulations might still be useful and should not be disregarded. We are currently exploring ways to accommodate these formulations. This mainly implies to refine the results of the web search stage, based on a broad statistical model of the context in which expressions of the original relationship are likely to occur. This should ensure that all the retrieved documents that are thematically close to the ones from which the seed sentence was originally extracted, and that the ‘general formulations’ they contain are nonetheless relevant.

The choice of the positive seed examples finally appears to be very important for the process of reformulation learning. Seed examples are more or less prototypical, and in some cases, it may be necessary to initialise the learning process with more than just one positive example to improve the initial set of formulations.

To summarise, these pilot studies have allowed us to validate the overall architecture of the learning machinery, and to get a much clearer view of the difficulty of the task at hand. They have also comforted our initial intuition that this mechanism had in fact the potential for extracting, in an unsupervised fashion, valid reformulations such as for instance "assassinate(x,y)" as a synonym of "kill(x,y)".

6. Conclusions and future work

In this paper, we have presented a general methodology for learning reformulations automatically in a nearly unsupervised fashion. The originality of this methodology is that it utilises the infrastructure NLP tools of an existing question answering system and that it searches for new formulations directly on the Web. The method proposed relies on a bootstrapping mechanism that alternatively acquires formulations and argument tuples. This bootstrapping mechanism is strongly inspired from recent work in information extraction, and has the potential for inducing simultaneously paraphrases and semantic classes.

The infrastructure of the learning system is now in place, and based on results of a pilot study, various improvements are currently under investigation, aiming at improving both the acquisition and the validation steps. The acquisition of new formulations will be completed by taking into account more complex lexico-syntactic structures. We will also try to take into account contextual elements before extracting arguments or formulations: the extraction of arguments and formulations will be done only if some contextual words are found in the documents. By increasing the number of documents analysed, the redundancy of the linguistic data found on the Web should become more obvious. The validation step will also be completed in various ways. In particular we intend to refine our statistical model, and to consider filtering the argument tuples according to their semantic categories, using an existing semantic network. We would finally like to extend our reformulation learning approach to other languages, such as English.

A last line of research that we need to develop concerns the evaluation of this methodology. To this end, we first have to develop a reasonable test set for the QA system and then measure the improvements incurred by the reformulation learning module.

7. References

Akira T., T. Takenobu (2001). Automatic disabbreviation by using context information. In Proceedings of the NLPRS 2002 Workshop on Automatic Paraphrasing: Theories and Applications.

Barzilay, R., K.R. McKeown (2001). Extracting paraphrases from a parallel corpus. In Proceedings of the Association for Computational Linguistics.

Collins, M., Y. Singer (1999). Unsupervised models for named entity classification. In Proceedings of the Workshop on Empirical Methods for Natural Language Processing EMNLP-VLC.

Brin, S. (1998). Extracting patterns and relations from the world wide web. In Proceedings of the WebDB Workshop at EDBT`98.

Boyer, M., G. Lapalme (1985). Generating paraphrases from meaning-text semantic networks. Computational Intelligence, 1(3&4),103--117.

Dras, M. (1998). Search in constraint-based paraphrasing. In Proceedings of the 2nd Conference on Natural Language Processing and Industrial Applications.

Duclaye, F., P. Filoche, J. Sitko, O. Collin (2002). A Polish question-answering system for business information. In Proceedings of the 5th International Conference on Business Information Systems (to appear).

Fuchs, C., 1982. La Paraphrase. Linguistique Nouvelle, Presses Universitaires de France.

Habert, B., A. Nazarenko, A. Salem (1997). Les linguistiques de corpus. In Armand Colin (Eds).

Hofman, T., J. Puzicha (1998). Statistical Models for co-occurrence data. MIT, AI Laboratory, Memo No.1625, C.B.C.L. Memo No.159.

Kurohashi, S., Y. 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, 481--488.

Martin R. (1976). Inférence, antonymie et paraphrase – Eléments pour une théorie sémantique. Collection Bibliothèque franç
aise et romane.

Mitamura, T., E. Nyberg (2001). Automatic rewriting for controlled language translation. In Proceedings of the NLPRS 2002 Workshop on Automatic Paraphrasing: Theories and Applications.

Murata, M., I. Hitoshi (2001). Universal model for paraphrasing – Using transformation-based on a defined criteria. In Proceedings of the NLPRS 2002 Workshop on Automatic Paraphrasing: Theories and Applications.

Ohtake, K., Y. Kazuhide (2001). Paraphrasing honorifics. In Proceedings of the NLPRS 2002 Workshop on Automatic Paraphrasing: Theories and Applications.

Riloff, E., R. Jones (1999). Learning dictionaries for information extraction by multi-level bootstrapping. In Proceedings of the 16th National Conference on Artificial Intelligence.

Sekine S. (2001). Extracting synonymous expressions from multiple newspaper documents. In Proceedings of the ANLP Workshop on Automatic Paraphrasing.

Tomuro N., S.L. Lytinen (2001). Selecting features for paraphrasing question Sentences. In Proceedings of the NLPRS 2002 Workshop on Automatic Paraphrasing: Theories and Applications.

Torisawa, K. (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.

Pirrelli,V., F.Yvon, (1999). The hidden dimension: a paradigmatic view of data-driven natural language processing. In Journal of Experimental and Theoretical Artificial Intelligence, 11, 391--408.