Automatic lexical acquisition from corpora: some limitations and tentative solutions

Cédric Messiant, Thierry Poibeau

Automatic Lexical Acquisition from Corpora. Some Limitations and some Tentative Solutions

Cédric Messiant[1] and Thierry Poibeau[2]
Laboratoire d'Informatique de Paris-Nord, Laboratoire LaTTiCe

Abstract

This paper deals with lexical acquisition. We take another look at some experiments we have recently carried out on the automatic acquisition of lexical resources from French corpora. We describe the architecture of our system for lexical acquisition. We formulate the hypothesis that some of the limitations of the current system are mainly due to a poor representation of the constraints used. Finally, we show how a better representation of constraints would yield better results.

Keywords: lexical acquisition from corpora; syntactic lexicon; subcategorization frames

1 Introduction

Natural Language Processing (NLP) aims at developing techniques to process natural language texts using computers. In order to yield accurate results, NLP requires voluminous resources containing various information (e.g. subcategorization frames—thereafter SCF, semantic roles, restriction of selection, *etc*.). Unfortunately, such resources are not available for most languages and they are very costly to develop manually. This is the reason why a lot of recent research has been devoted to the automatic acquisition of resources from corpora.

Automatic lexical acquisition is an engineering task aiming at providing comprehensive—even if not fully accurate— resources for NLP. As natural languages are complex, lexical acquisition needs to take into account a wide range of parameters and constraints (cf. mainly the kind of information detailed in the previous paragraph along with frequency information) However, surprisingly, in the acquisition community, relatively few investigations have been conducted on the structure of the linguistic constraints themselves.

In this paper, we want to take another look at some experiments we have recently carried out on the automatic acquisition of lexical resources from French corpora. The task consists, from a surface form, in trying to find an abstract lexical-conceptual structure that justifies the surface construction (taking into account the relevant set of constraints for the given language). Here, in order to get a tractable model, we limit ourselves to the acquisition of subcategorization frames from corpora. The task is

[1] CNRS and Université Paris 13, cedric.messiant@lipn.univ-paris13.fr

[2] CNRS and Ecole Normale Supérieure, thierry.poibeau@ens.fr

challenging since surface form incorporates adverbs, modifiers, interpolated clauses and some flexibility in the order of appearance of the arguments that, of course, should not affect the analysis of the underlying lexical-conceptual structure.

Most approaches, including ours, are based on simple filtering techniques. If a complement appears very rarely associated with a given predicate, the acquisition process will assume that this is an incidental co-occurrence that should be left out. However, as we will see, even if this technique is efficient for high frequency items, it leaves a lot of phenomena aside.

2 Previous Work in Lexical acquisition from corpora

Large corpora and efficient parsers are now widely available for a growing number of languages. So, even though lexical resources are not always available, it is now possible to acquire large lexicons directly from the observation of word usage in corpora, based on the output of surface parsers. Moreover, using automatic acquisition techniques makes it possible to get frequency information associated with lexical entries, which is not possible simply using a manual approach.

Several systems have been built using this approach, for several languages — see, among others, Brent (1993), Manning (1993), Briscoe and Carroll (1997), Korhonen (2002), Schulte im Walde (2002), Messiant (2008) and Messiant *et al*. (2008). The acquisition process is made of three different steps:

1. all the occurrences of the different verbs are grouped together, along with their complements;
2. tentative constructions for each verb are identified, along with their respective productivity (we call these "tentative constructions" since they may contain modifiers, and not only arguments; tentative constructions need to be filter to give birth to actual subcategorization frames);
3. rare constructions are filtered out, taking as an hypothesis the fact that too few occurrences of a construction is probably the sign of an error in the analysis (or a sign that the construction includes an adjunct).

All the systems are based on these hypotheses, even though they differ as for their parsing model or filtering strategy.

3 A Lexical Acquisition System for French

3.1. Pre-processing: Morpho-syntactic tagging and syntactic analysis

Our system first tags and lemmatizes corpus data using the TreeTagger and then parses it using Syntex (Bourigault *et al*., 2005). Syntex is a shallow parser for French. It uses a combination of heuristics and statistics to find dependency relations between tokens

in a sentence. It is a relatively accurate parser, *e.g.* it obtained the best precision and F-measure for written French text in the recent EASY evaluation campaign[3].

Below is an example that illustrates the dependency relations detected by Syntex (2) for the input sentence in (1):

(1) *La sécheresse s'abattit sur le Sahel en 1972-1973* .
(*The drought came down on Sahel in 1972-1973.*)

(2)
```
DetFS|le|La|1|DET;2|
NomFS|sécheresse|sécheresse|2|SUJ;4|DET;1
Pro|se|s'|3|REF;4|
VCONJS|abattre|abattit|4|SUJ;2,REF;3,PREP;5,PREP;8
Prep|sur|sur|5|PREP;4|NOMPREP;7
DetMS|le|le|6|DET;7|
NomMS|sahel|Sahel|7|NOMPREP;5|DET;6
Prep|en|en|8|PREP;4|NOMPREP;9
NomXXDate|1972-1973|1972-1973|9|NOMPREP;8|
Typo|.|.|10||
```

Syntex does not make a distinction between arguments and adjuncts—rather, each dependency of a verb is attached to the verb.

3.2 Pattern extractor

The pattern extractor collects the dependencies found by the parser for each occurrence of a target verb. Some cases receive special treatment in this module. For example, if the reflexive pronoun "se" is one of the dependencies of a verb, the system considers this verb like a new one. In (1), the pattern will correspond to "s'abattre" and not to "abattre". If a preposition is the head of one of the dependencies, the module explores the syntactic analysis to find if it is followed by a noun phrase (`+SN`) or an infinitive verb (`+SINF`).

Example (3) shows the output of the pattern extractor for the input in (1).

(3)
```
VCONJS|s'abattre :
Prep+SN|sur|PREP Prep+SN|en|PREP
```

3.3 The Subcategorization Frame builder

The SCF builder extracts SCF candidates for each verb from the output of the pattern extractor and calculates the number of corpus occurrences for each SCF and verb combination. The syntactic constituents used for building the SCFs are the following:

1. `SN` for nominal phrases;
2. `SINF` for infinitive clauses;

[3] The scores and ranks of Syntex at this evaluation campaign are available at http://w3.univ-tlse2.fr/erss/textes/pagespersos/bourigault/syntex.html#easy

3. `SP[prep+SN]` for prepositional phrases where the preposition is followed by a noun phrase (prep is the prepositional head);
4. `SP[prep+SINF]` for prepositional phrases where the preposition is followed by an infinitive verb (prep is the prepositional head);
5. `SA` for adjectival phrases;
6. `COMPL` for subordinate clauses.

When a verb has no dependency, its SCF is considered as `INTRANS`.

Example (4) shows the output of the SCF builder for (1).

(4) `S'ABATTRE+s'abattre ;;; SP[sur+SN] SP[en+SN]`

3.4 The Subcategorization Frame Filter

Each step of the process is fully automatic, so the output of the SCF builder is noisy due to tagging, parsing or other processing errors. It is also noisy because of the difficulty of the argument-adjunct distinction. The latter is difficult even for humans. Many criteria are not usable because they either depend on lexical information which the parser cannot make use of (since our task is to acquire this information) or on semantic information which even the best parsers cannot yet learn reliably. Our approach is based on the assumption that true arguments tend to occur more regularly and more frequently after the verb than adjuncts. Thus many frequent SCFs in the system output are correct.

We therefore filter low frequency entries from the SCF builder output. We currently do this using Maximum Likehood Estimates (Korhonen, Gorrell, & McCarthy, 2000). This simple method involves calculating the relative frequency of each SCF (for a verb) and comparing it to an empirically determined threshold. The relative frequency of the SCF *i* with the verb *j* is calculated as follows:

$$rel\_freq(scf_i, verb_i) = \frac{|scf_i, verb_j|}{|verb_j|}$$

$|scf_i, verb_j|$ is the number of occurrences of the SCF *i* with the verb *j* and $|verb_j|$ is the total number of occurrences of the verb *j* in the corpus.

If, for example, the frequency of the SCF `SP[sur+SN] SP[en+SN]` is less than the empirically defined threshold, the SCF is rejected by the filter. The Maximum Likehood Estimates filter is not perfect because it is based on rejecting low frequency SCFs, which leads to sometimes reject frames that are indeed correct. Our filter incorporates specific heuristics for cases where this assumption tends to generate too many errors. With prepositional SCFs involving one prepositional phrase (PP) or more, the filter determines which one is the less frequent PP. It then re-assigns the associated frequency to the same SCF without this PP.

For example, `SP[sur+SN] SP[en+SN]` could be split into two SCFs : `SP[sur+SN]` and `SP[en+SN]`. In our example, `SP[en+SN]` is the less frequent prepositional phrase and the final SCF for the sentence (1) is (5).

(5) `SP[sur+SN]`

Note that `SP[en+SN]` is here an adjunct.

4 Some difficulties with this kind of approach

This approach is very efficient to deal with large corpora. However, some issues remain. As the approach is based on automatic tools (especially parsers) that are far from perfect, the obtained resources always contain errors and have to be manually validated. Moreover, the system needs to get sufficient examples to be able to infer relevant information. Therefore, there is generally a lack of information for a lot of low productivity items (the famous "*sparse problem*").

More fundamentally, some constructions are difficult to acquire and characterise automatically. On the one hand, idioms are not recognised as such by most acquisition systems. On the other hand, some adjuncts appear frequently with certain verbs (*eg*. verbs like *dormir* 'to sleep' frequently appear with location complements). The system then assumes that these complements are arguments, whereas linguistic theory would say without any doubt that these are adjuncts. Lastly, surface cues are sometimes insufficient to recognize ambiguous constructions (cf. …*manger une glace à la vanille* 'to eat a vanilla ice-cream' vs *manger une glace à la terrasse d'un café* 'to eat an ice-cream outside the café').

5 Some solutions

These issues do not mean that automatic methods are flawed, but that they have a number of drawbacks that should be addressed. The acquisition process, based on an analysis of co-occurrences of the verb with its immediate complements (along with filtering techniques), makes the approach highly functional. It is a good approximation of the problem. However, this model does not take into account external constraints.

5.1. Idioms and light verb constructions

The fact that some phrasal complements (with a specific head noun) frequently co-occur with a given verb is most of the time useful, especially to identify idioms (Fabre and Bourigault, 2008), colligations (Firth, 1968) and light verb constructions (Butt, 2003). On the other hand, the fact that a given prepositional phrase appears with a large number of verbs may indicate that the preposition introduces an adjunct rather than an argument.

So, instead of simply capturing the co-occurrences of a verb with its complements, we have a number of important features which are available:

- indicator of the dispersion of the prepositional phrases (PPs) depending on the prepositional head (if a PP with a given preposition appears with a wide range of different verbs, it is more likely to be a modifier);
- indicator of the probability for a given PP to appear as an argument rather than as an adjunct (some PPs are rarely arguments, e.g. time or location phrases);
- indicator of the co-occurrence of the nominal head of an argument (NP or PP) with a verb (if a verb appears frequently with the same nominal head, it is more likely to form a semi-idiomatic expression);
- indicator of the complexity of the sentence to be processed (if a sentence is complex, its analysis is less reliable). We can calculate a "confidence measure" of the syntactic analysis of a sentence and thus of the syntactic frame extracted from this sentence;
- lastly, semantic typing of the arguments, to distinguish two similar SCFs if they differ only from a semantic point of view.

To be able to do this, the pattern extractor has to be modified in order to keep most of the information that was previously rejected as not relevant. We then need to calculate these indicators so that they can be taken into account.

All these constraints can be evaluated separately, so that we obtain for each of them an ideal evaluation of the parameter. There are two ways of doing this:

1) by automatically inferring the different weights from a set of annotated data

or

2) by estimating the results of various manually defined weights.

We are currently using the second method since data annotation is very costly. However, the first approach would certainly lead to more accurate results. The weight and the ranking of the different constraints must then be examined. A linear model can provide a first approximation but there are surely better ways to integrate the different constraints. Some studies may provide some cues but we still need to evaluate them in our framework (Blache and Prost, 2008).

This is the reason why we are interested in constraints models. We assume that language can be represented using a set of constraints, themselves modelled as "dynamic forces". The same idea has been developed in various theories (e.g. Shieber 1992 ; Blache 2001). However, it seems that it has not been fully developed in the case of acquisition processes.

5.2 Manual Validation

The approach requires manual validation. Rather than leaving the validation process apart for further tedious examination by a linguist, we propose to integrate it in the acquisition process itself. Taking into consideration the number of examples and the complexity of the sentences used for training, it is possible to associate confidence scores with the different constructions of a given verb: the linguist is then able to quickly focus on the most problematic cases. It is also possible to propose tentative constructions to the linguist, when not enough occurrences are available for training.

Lastly, when too few examples are available, the linguist can provide relevant information to the machine. However, a well-designed and dynamic validation process makes it possible to decrease by one order of magnitude the time spent on validating the data (Figure 1 presents an overview of the system interface).

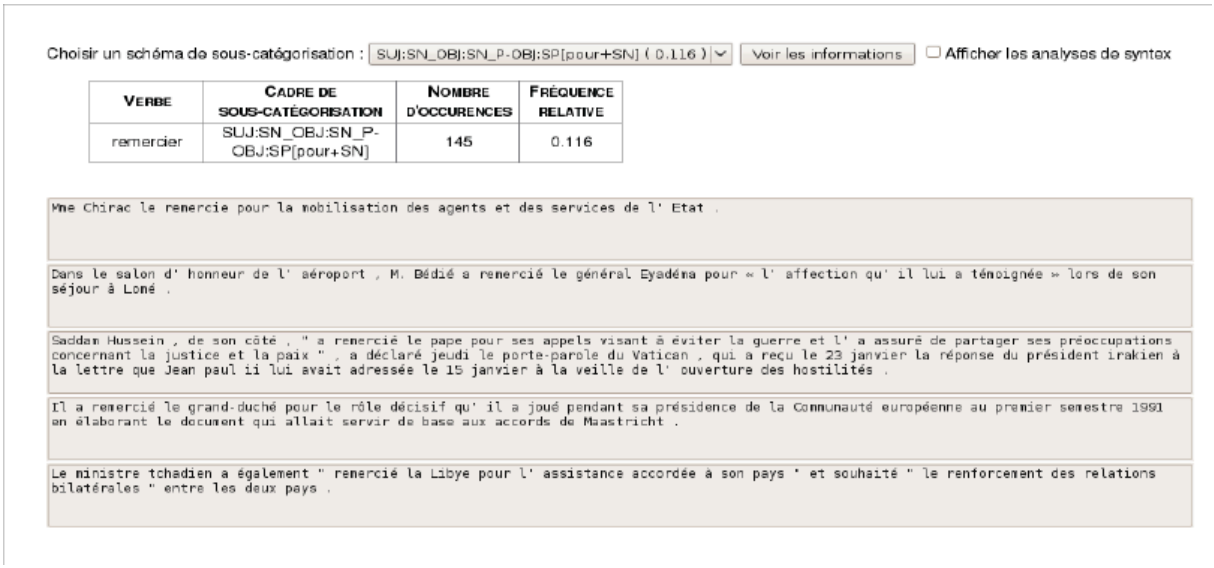

Figure 1: An overview of the interface of the system: http://www-lipn.univ-paris13.fr/~messiant/lexschem.html

6 Conclusion

This paper introduced LexSchem – a large-scale subcategorization frame lexicon for French verbs. The lexicon has been automatically acquired from a large corpus and currently contains 10,928 lexical entries for 5,261 French verbs. The lexicon is provided with a graphical interface and is made freely available to the community via a web page. Future work will include improvement of the filtering module (e.g. experimenting with SCF-specific thresholds or smoothing using semantic back-off estimates), automatic acquisition of subcategorization frames for other French word classes (e.g. nouns), and automatic classification of verbs using the subcategorization frames as features (Levin, 1993).

Acknowledgements

This research was carried out as part of an Alliance grant funded by the British Council and the French Ministry of Foreign Affairs; Cédric Messiant's PhD is funded by a DGA/CNRS Grant.

References

BLACHE, P. (2001). *Les Grammaires de Propriétés: des contraintes pour le traitement automatique des langues naturelles.* Paris: Hermes Sciences.

BLACHE, P. and PROST, J.-P. (2008). A Quantification Model of Grammaticality. In *Proceedings of the 5th International Workshop on Constraints and Language Processing* (CSLP2008). Hamburg.

BRENT, M.R. (1993). From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax. *Computational Linguistics*, 19: 203–222.

BOURIGAULT D., JACQUES M.-P., FABRE C., FRÉROT C., OZDOWSKA S., (2005). Syntex, analyseur syntaxique de corpus. In *Actes de la conférence Traitement Automatique des Langues Naturelles* (TALN), Dourdan.

BUTT, M. (2003). The Light Verb Jungle. *Harvard Working Papers in Linguistics*, 9: 1-49.

BRISCOE, T. and CARROLL, J. (1997). Automatic extraction of subcategorization from corpora. *Proceedings of the Meeting of the Association for Computational Linguistics*. Washington: 356–363.

FABRE, C. and BOURIGAULT, D. (2008). Exploiter des corpus annotés syntaxiquement pour observer le continuum entre arguments et circonstants. J*ournal of French Language Studies*. 18(1): 87–102.

FIRTH, J.R. (1968). A synopsis of linguistic theory. In F.R. Palmer (ed.). *Selected Papers of J.R. Firth 1952–59*. London: Longmans, 168–205.

KORHONEN, A. (2002). Subcategorization Acquisition. *Technical Report UCAM-CL-TR-530*. University de Cambridge: Computer Laboratory.

KORHONEN A., GORRELL G., MCCARTHY D. (2000). Statistical filtering and subcategorization frame acquisition. *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, Hong Kong.

LEVIN B. (1993). English Verb Classes and Alternations: a preliminary investigation. Chicago: University of Chicago Press.

MANNING, C.D. (1993). Automatic Acquisition of a Large Subcategorization Dictionary from Corpora. *Proceedings of the Meeting of the Association for Computational Linguistics*. Columbus: 235–242.

MESSIANT, C. (2008). ASSCI: A Subcategorization Frames Acquisition System for French Verbs. *Proceedings of the Conference of the Association for Computational Linguistics* (ACL, Student Research Workshop), Columbus: 55–60.

MESSIANT, C., KORHONEN, A. and POIBEAU, T. (2008). LexSchem: A Large Subcategorization Lexicon for French Verbs. *Proceedings Language Resources and Evaluation Conference* (LREC). Marrakech.

POIBEAU, T. and MESSIANT, C. (2008). Do we still Need Gold Standards for Evaluation? . *Proceedings Language Resources and Evaluation Conference* (LREC). Marrakech.

PRINCE, A. and SMOLENSKY, P. (2004). Optimality Theory: Constraint Interaction in Generative Grammar. Oxford: Blackwell.

SCHULTE IM WALDE, S. (2002). A Subcategorisation Lexicon for German Verbs induced from a Lexicalised PCFG. *Proceedings Language Resources and Evaluation Conference* (LREC). Las Palmas: 1351–1357.

SHIEBER, S. (1992). *Constraint-based Grammar Formalisms*. Cambridge: The MIT Press.