tripleC 4(2): 277-283, 2006
ISSN 1726-670X
http://tripleC.uti.at

Symbolic Machine Learning: A Different Answer to the Problem of the Acquisition of Lexical Knowledge from Corpora

Pascale Sébillot

IRISA, Campus de Beaulieu,
35042 Rennes cedex, France,
sebillot@irisa.fr,
33 2 99 84 73 17

Abstract: One relevant way to structure the domain of lexical knowledge (*e.g.* relations between lexical units) acquisition from corpora is to oppose numerical *versus* symbolic techniques. Numerical approaches of acquisition exploit the frequential aspect of data, have been widely used, and produce portable systems, but poor explanations of their results. Symbolic approaches exploit the structural aspect of data. Among them, the symbolic machine learning (ML) techniques can infer efficient and expressive patterns of a target relation from examples of elements that verify this relation. These methods are however far less known, and the aim of this paper is to point out their interest through the description of one precise experiment. To remove their supervised characteristic, and instead of opposing them to numerical approaches, we finally show that it is possible to combine one symbolic ML technique to one numerical one, and keep advantages of both (meaningful patterns, efficient extraction, portability).

Keywords: corpus-based lexical knowledge acquisition, symbolic machine learning, inductive logic programming, numerical approaches, semi-supervised machine learning.

Acknowledgement: A large part of the work reported in this paper is due to a deep collaboration with Vincent Claveau (IRISA).

1 Introduction

The performances of a lot of natural language applications (automatic translation, information retrieval, *etc.*) rely on the quality of their lexical semantic resources. As the contents of those resources vary from one domain and one application to another, a common solution to have them at disposal is to automatically extract them from textual corpora; a lot of studies have been dedicated to this purpose.

One relevant way to structure this domain of lexical knowledge (complex terms, or relations between lexical units) acquisition from corpora is to oppose numerical *versus* symbolic techniques. Numerical approaches of acquisition exploit the frequential aspect of data, and use statistical techniques, while symbolic approaches exploit the structural aspect of data, and use structural or symbolic information[1]. Those two families of techniques are thus different answers to a same problem, with distinct strong points

[1] Note that no assumption is made about the actual technique manipulating symbolic or numerical information; a statistical technique can be used to acquire terms or lexical relations on the basis of symbolic information, and conversely, a symbolic technique can make the most of numerical information. In this paper, the words methods, techniques and approaches are indifferently used as synonyms for either numerical or symbolic approaches of acquisition.

and drawbacks. But if the first kind of methods has been widely employed both to extract complex terms or syntagmatic, and paradigmatic relations (Manning and Schütze 1999), only one of the two strategies of the symbolic approach of acquisition has been really investigated: the symbolic linguistic approach, in which patterns of the elements to acquire are manually established by linguists (*e.g.* (Oueslati 1999)). When such patterns are unknown, the second strategy of the symbolic approach, *i.e.*, symbolic machine learning (ML) (Mitchell 1997) proposes solutions to automatically learn them from examples of elements respecting the target relations or form of the terms. Indeed, this technique offers in particular answers for cases in which other approaches, especially numerical ones, cannot be used. This second facet of the symbolic approach, far less known and employed, is just beginning to appear and widen in the natural language processing community.

Our aim here is therefore to point out the interest of the ML approach of acquisition. In order to demonstrate what those techniques can provide, this paper, through the description of an experiment concerning the acquisition of patterns of one type of semantic relations, explains how one ML method, inductive logic programming (ILP) (Muggleton and De Raedt 1994), works, and what limits of numerical approaches such a technique can solve. After a first section dedicated to the key-points of numerical approaches, including their advantages and drawbacks, and to a general presentation of the symbolic approach, we describe the ILP experiment, and weak and strong elements of the technique. Rather than an opposition between those two approaches, the concluding section explains how they can collaborate, to hold all their advantages.

2 Numerical *versus* symbolic approach of acquisition

As mentioned above, one relevant way to structure the domain of lexical relation acquisition from corpora is to oppose numerical *versus* symbolic approaches. The former approach of acquisition exploits the frequential aspect of data while the latter exploits the structural aspect of data, and uses symbolic information.

Within the numerical approach, complex terms or relations between lexical units can be acquired by studying word co-occurrences in a text window (or specific syntactic structures). The strength of the association is usually evaluated with the help of a statistical score (association coefficient; *e.g.*, mutual information, loglike, *etc.*) that detects words appearing together in a statistically significant way. For example, Kenneth W. Church and Patrick Hanks's work (1989) is based on such a statistical co-occurrence method. Following Harris's linguistic principles (Harris *et al.* 1989), numerical distributional analysis methods respect a 3-step approach: extraction of the co-occurrents of one word (within a text window or a syntactic context), evaluation of proximity/distance between two terms, based on their shared or not shared co-occurrents (various measures are defined), clustering into classes (following different data analysis or graph techniques). For example, Jacques Bouaud *et al.* (1997) and Gregory Grefenstette (1994) use this kind of technique to discover paradigmatic relations.

Let us briefly sum up the advantages and drawbacks of numerical methods: they are portable, robust, and automatic but produce non-interpretable results; the detection is realized at the corpus level: thus, the detection of one specific occurrence cannot be explained; and rare cases may be problematic. This non-interpretability of results may become a problem, especially in the domain of computational linguistics, when discussing them with a linguist. Complex terms or examples of pairs of words respecting a given relation cannot be linguistically justified, and no interesting contrastive discussion of the characteristics of accepted or non-accepted elements can be done. It may thus be difficult to consider the sets of complex terms or of pairs of related words obtained as *true* knowledge because abstracting their properties and specificities is very difficult.

The symbolic approach of acquisition groups two strategies: the symbolic linguistic approach, and the machine-learning (ML) approach. In the first one, operational definitions of the elements to acquire are manually established by linguists, usually in the form of morpho-lexical patterns that carry the relations

that are studied, or by a list of linguistic clues (*e.g.* (Oueslati 1999)). However, when such patterns or clues are unknown[2], but examples of elements respecting the target terms or relation are known, ML can be used to automatically extract patterns from the descriptions of those examples. The technique is based on a 5-step methodology initiated by Marti A. Hearst (1992):

1. select one target relation R;
2. gather a list of pairs following relation R;
3. find the sentences that contain those pairs; keep their lexical and syntactic contexts;
4. detect common points between those contexts; suppose that they form a pattern for R;
5. apply the patterns to get new pairs and go back to 3.

Symbolic ML (inductive logic programming, grammatical inference, *etc.*) (Mitchell 1997) offers a framework to automate step 4, and aims at automatically producing unknown morpho-lexical patterns that carry the target terms or relation.

3 ILP to acquire Noun-Verb relations

The use of symbolic ML methods is just beginning to widen in the natural language processing community. Among these methods, ILP (Muggleton and De Raedt 1994), thanks to its expressiveness and flexibility, has been applied to different problems (overview in (Cussens and Džeroski 2000)). In this section we first summarize the main principles of ILP; we then explain the application of our ILP system (Claveau *et al.* 2003), ASARES, to the acquisition of extraction patterns for some specific semantic relations between a noun and a verb. We finally present and discuss the results of this experiment and particularly stress their interpretability and their linguistic interest.

3.1 Principles of ILP

ILP aims at producing general rules (more precisely Horn clauses) explaining a concept from examples and counter-examples of the concept and from a background knowledge.

A hypothesis language is also provided to the ILP system; it is used to precisely define the expected form of the generated rules (or hypotheses). According to this language, the ILP algorithm infers rules that cover (that is, explain, characterize) a maximum of examples and no counter-examples (or only a few, some *noise* can be allowed in order to produce more general patterns), by generalizing the examples in a controlled way. More precisely, the inference process conforms to the following steps:

1. select one example *e* in the set of examples to be generalized. If none exists, stop;
2. define a hypothesis search space *H* according to *e* and the hypothesis language;
3. search *H* for the rule *h* that maximizes a score function *Sc*;
4. remove the examples that are covered by the chosen rule. Return to step 1.

The score function *Sc* depends on the number of examples and counter-examples covered by a hypothesis *h*.

The result of the inference process is the set of hypotheses *h* that is obtained and that corresponds to interpretable patterns of the target concept.

3.2 Acquisition of Noun-Verb relations

Though not dedicated to any particular semantic link, we have applied our ILP system (Claveau *et al.* 2003) ASARES, in an illustrative aim, to the acquisition of extraction patterns for some semantic relations between a noun (N) and a verb (V): the inferred rules or patterns must allow us to extract N-V couples in which V plays one of the *qualia* roles of N, as defined in the Generative Lexicon (GL) formalism (Pustejovsky 1995). GL is a lexicon model in which lexical entries consist of structured sets of predicates

[2] Or are domain-dependent.

that define a word. In one of the components of this model, called the *qualia structure*, words are described with semantic roles such as the purpose or function (*e.g.*, cut for knife), or the creation mode (build for house), *etc.* For a given word, each role can get numerous realizations, and the qualia structure of each word, especially for common nouns, is mainly made up of verbal associations. Such N-V pairs, in which V plays one of the qualia roles (function, creation mode, *etc.*) of N, are called *qualia pairs* hereafter.

Thus, in our ILP framework, the concept to be learned is the qualia nature of a N-V pair occurring within a sentence. Using a ML technique is especially well-suited here both because the extraction patterns for qualia relations are not known, and statistical co-occurrence-based methods have been proved not satisfactory for this task (see (Bouillon *et al.* 2002) and section 3.3). An example (resp. counter-example) corresponds to a N-V couple manually indicated by an expert as verifying (resp. not verifying) the target qualia relations in one sentence of a our corpus. For example, the pair (screwdriver-screw) in its context: "*The operator uses the screwdrivers to screw…*"[3] can be used to code an example, whereas (tyre-prescribe) in its context: "*Inflate the tyre to the prescribed pressure...* " can form a counter-example. For our experiment, the corpus used is a collection of helicopter maintenance handbooks, provided to us by MATRA-CCR Aérospatiale. This French 700 KBytes technical corpus contains more than 104,000 word occurrences, and has been Part-of-Speech and semantically tagged (see (Bouillon *et al.* 2002) for details concerning this second tagging). Hierarchies of PoS and semantic tags are parts of the knowledge inserted in the background knowledge (for generalizations).

A GL expert has thus extracted from the corpus qualia and non-qualia N-V pairs with their contexts (all the words and their tags occurring with the pair within a sentence). ASARES, our ILP system, has been given this way about 3,000 examples and 3,000 counter-examples.

The ILP system automatically infers rules, *i.e.*, extraction patterns for the target relations like: *is_qualia(N,V) :- precedes(V,N), near_verb(N,V), infinitive(V), action_verb(V).* which means that a pair composed by a noun N and a verb V will be considered as qualia if N appears in a sentence after V, N and V are not far from each other (especially not separated by a verb), and V is an action verb in the infinitive.

The nine rules obtained are given below:

1. is_qualia(N,V) :- precedes(V,N), near_verb(N,V), infinitive(V), action_verb(V).
2. is_qualia(N,V) :- contiguous(N,V).
3. is_qualia(N,V) :- precedes(V,N), near_word(N,V), near_verb(N,V), suc(V,X), preposition(X).
4. is_qualia(N,V) :- near_word(N,V), sentence_beginning(N).
5. is_qualia(N,V) :- precedes(N,V), singular_common_noun(N), suc(V,C), colon(C), pred(N,D), punctuation(D).
6. is_qualia(N,V) :- near_word(N,V), suc(V,C), suc(C,D), action_verb(D).
7. is_qualia(N,V) :- precedes(N,V), near_word(N,V), pred(N,C), punctuation(C).
8. is_qualia(N,V) :- near_verb(N,V), pred(V,C), pred(C,D), pred(D,E), preposition(E), sentence_beginning(N).
9. is_qualia(N,V) :- precedes(N,V), near_verb(N,V), pred(N,C), subordinating_conjunction(C).

Those inferred rules are then used as qualia N-V pair extraction patterns to retrieve new qualia pairs from the corpus.

3.3 Results and discussions

As explained in (Bouillon *et al.* 2002), using the produced patterns to extract qualia N-V pairs from the corpus gives good results. The evaluation has been conducted on a test-set formed by N-V pairs in which N is one of seven domain relevant common nouns: (screw, nut, door, indicator signal, plug, cowl, cap)[4].

[3] For understanding reasons, all the examples given in this paper are translations of those obtained from our French corpus.

[4] To prevent distortion of results, none of these common nouns were used as examples or counter-examples for the pattern induction in the ILP system.

Each occurrence of this kind of pairs in a part (32,000 words) of our corpus is manually annotated as qualia or not qualia by GL experts, and the extraction results of patterns produced by our ILP system concerning the same pairs are compared with the human experts's decisions. This paper (Bouillon *et al.* 2002) also stresses one crucial result concerning lexical knowledge acquisition: for very precise relations like qualia ones, numerical techniques are far less relevant[5] than our symbolic approach (the 2 best results obtained are given together with those of our ILP method in Table 1 below; in the numerical framework, a qualia pair is considered as a special kind of co-occurrence, and the strength of the association is measured by an association coefficient).

System	Precision (P)	Recall (R)	F-measure[6]
ILP (ASARES)	62.2%	92.4%	0.744
Ochiai coefficient	82.4%	42.4%	0.56
MI^3 coefficient[7]	92.3%	36.4%	0.522

Table 1: Comparisons of ILP-based and statistical method results

As previously mentioned, another interest of symbolic learning techniques like ILP is the production of meaningful patterns; through the produced rules, ILP gives here access to a linguistically interpretable support to the concept of qualia role. We can thus examine the linguistic relevance of the nine generated extraction patterns.

What is first striking is the fact that, at the level of generalization reached here, few linguistic features are retained. The clauses seem to provide very general indications and tell us very little about types of verbs (action verb is the only information we get), nouns (common noun) or prepositions that are likely to fit into such structures. But the clauses contain other information, related to several aspects of linguistic descriptions, like:

- proximity: this is a major criterion; most clauses indicate that the noun and the verb must be either contiguous or separated by at most one element, and that no verb must appear between them;
- position: some clauses indicate that one of the two elements is found at the beginning of a sentence or right after a punctuation mark, whereas the relative position of N and V is given in others;
- punctuation: punctuation marks, and more specifically colons, are often mentioned. This kind of surface clues, very important here, are generally neglected by manual analysis;
- morpho-syntactic categorization: the first clause detects a very important structure in the text, corresponding to action verbs in the infinitive form.

These features bring to light linguistic patterns that are very specific to the corpus, a text falling within the instructional genre. We find in this text many examples in which a verb at the infinitive form occurs at the beginning of a proposition and is followed by a noun phrase. Such lists of instructions are very typical of the corpus (*e.g.*, disconnect the plug). Rule 5, which is equivalent to the pattern *V + : + (any token)* + [:,;] + singular N*, highlights enumerative structures that are very frequent in the corpus (*e.g.*, Open: the sliding cowl, the right cowl…). These results emphasize the ability of the ML technique to learn corpus-specific patterns.

After this rapid presentation of this ILP experiment, we can now summarize the weak and strong points of symbolic approaches: they need *a priori* knowledge (*e.g.* examples for ILP), but produce interpretable results; detection is done at the occurrence level, and rare cases can be treated. The knowledge extracted with the help of this ML approach is thus quite different from what is obtained with numerical techniques and may lead to interesting linguistic discussions.

[5] See (Bouillon et al. 2002) for a deep discussion concerning this point.
[6] F-measure= (2PR)/(P+R).
[7] Cubed mutual information coefficient.

4 Concluding remarks

Opposing numerical to symbolic approach of acquisition of lexical knowledge, after this presentation, indeed seems natural. Beyond the points already mentioned, some *rules* concerning the selection of one technique can even be proposed. Numerical approach is often very effective; but it may be problematic for very specific semantic relations. In that case, a symbolic approach has to be used; the same choice must be done when explanations of the results are needed. Of course, lots of other criteria influence the decision (size of the corpus, knowledge of *a priori* patterns, number of required examples, *etc.*).

But instead of opposing the two approaches, another interesting question concerns their possible combination in order to try and keep advantages from both families of techniques:

- quality of results, and interpretability of the supervised symbolic extraction;
- automaticity, and portability of the statistical extraction.

This type of combination is indeed possible and we conclude this paper by a rapid description of the production of so-called *semi-supervised* versions of ASARES.

In the previous description of our ILP method, the cost essentially lies in the construction by an expert of example and counter-example sets; this makes the technique time-consuming, and thus difficult to apply to a new corpus. However, we have shown in (Claveau and Sébillot 2004) that it is possible to combine one statistical co-occurrence-based method with our symbolic system in order to overcome this problem. Bootstrapping the ILP method by the numerical one leads to two combinations that preserve advantages of each of the different extraction approaches and rival the performances of the former *supervised* (*i.e.*, fed by examples) ILP system.

The first hybrid system proposed relies on a sequential combination of a statistical and the symbolic system. Here, the statistical system is a simple co-occurrence technique based on the MI^3 coefficient. Each system iteratively uses as input the output data of the other one. More precisely, the N-V pair list generated by a system is used by the other one to construct its own N-V pair list. The only constraint is to begin this iteration with the statistical system since it does not need any data but the corpus. The loop terminates when the same set of rules is obtained during two successive iterations. The resulting extraction technique is called hereafter sequential hybrid system.

Unlike this first system in which the statistical and ILP-based systems are used without major modifications, our second hybrid extraction technique combines them more finely and implies some changes in the ILP algorithm. As mentioned in section 3.1, during the third learning step, a rule *h* is chosen from a hypothesis space *H* if it maximizes a score function *Sc* that depends on the number of examples and counter-examples it covers. The principle of our second hybrid system is to weight the examples according to their statistical scores so that the hypotheses are now evaluated with the help of these weighted examples. The sets of weighted examples and counter-examples are thus built with the MI^3 extraction system: the highest MI^3 scored pairs are considered as examples, and conversely, the lowest as counter-examples; their weights *w* are computed from their MI^3 scores. Therefore, the more the example is considered interesting (that is, highly scored) by the statistical technique, the more it influences the choice of rules. Finally, the rules that are kept are those maximizing the score function *Sc* redefined as the sum of the weights of examples that it covers minus the sum of the weights of counter-examples that it covers. This extraction technique, which is less expensive (*i.e.*, it requires only one statistical extraction and one ILP learning phase), is called integrated hybrid system.

The two systems have been evaluated for the same task (acquisition of qualia extraction patterns) on the same test-set as the one described in section 3.3. The patterns obtained are quite similar to those inferred by the supervised version of ASARES. Table 2 below summarizes the performance of the three methods.

System	Precision (P)	Recall (R)	F-measure
Supervised ASARES	62.2%	92.4%	0.744
Sequential hybrid	62.0%	93.9%	0.747

Integrated hybrid	60.2%	89.4%	0.720

Table 2: Performances of the three systems

The resulting semi-supervised versions of ASARES thus rival its supervised version and fulfill our objectives: they produce efficient patterns, able to extract, once applied to a corpus, pairs of elements actually bound by the target relations; these patterns are expressive, that is, linguistically relevant; and, these methods are generic and, thanks to the use of the numerical technique, easily portable from one corpus/relation to another.

In this paper, our aim was to shed light on the symbolic machine learning approach of corpus-based acquisition of lexical knowledge, a family of techniques still rarely used in computational linguistics. The description of a precise experiment has allowed us to stress its main strong points: efficiency and interpretability of the inferred patterns that can lead to interesting linguistic discussions; possibility to treat fine-grained relations difficult to grasp with other kinds of approaches. The necessity of supervision of such ML techniques can be circumvented by combining this symbolic approach with a statistical method. Both ML approach, and semi-supervised systems combining ML and numerical approaches are promising tracks still to be deeper explored in the field of computational linguistics.

References

Bouaud J., Habert B., Nazarenko A. and Zweigenbaum P. (1997) *Regroupements issus de dépendances syntaxiques en corpus : catégorisation et confrontation avec deux modélisations conceptuelles*. In: Proceedings of Ingénierie des Connaissances IC'97. pp 207-223.

Bouillon P., Claveau V., Fabre C., Sébillot P. (2002) *Acquisition of Qualia Elements from Corpora - Evaluation of a Symbolic Learning Method*. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation LREC'02. pp 208-215.

Church K. W., Hanks P. (1989) *Word Association Norms, Mutual Information, and Lexicography*. In: Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics ACL'89. pp 76-83.

Claveau V., Sébillot, P. (2004) *From Efficiency to Portability: Acquisition of Semantic Relations by Semi-supervised Machine Learning*. In: Proceedings of the 20th International Conference on Computational Linguistics COLING'04. pp 261-267.

Claveau, V., Sébillot P., Fabre C., Bouillon P. (2003) *Learning Semantic Lexicons from a Part-of-Speech and Semantically Tagged Corpus using Inductive Logic Programming*. In: Journal of Machine Learning Research, special issue on Inductive Logic Programming. Vol. 4. pp 493-525.

Cussens J., Džeroski S. (2000) *Learning Language in Logic*. Vol. 1925. LNAI,Lecture Notes in Computer Science. Springer Verlag.

Grefenstette G. (1994) *Explorations in Automatic Thesaurus Discovery*. Dordrecht. Kluwer Academic Publishers.

Harris, Z., Gottfried M., Ryckman T., Mattick P.(Jr), Daladier A., Harris T. N., Harris S. (1989) *The Form of Information in Science. Analysis of Immunology Sublanguage*. In: Boston Studies in the Philosophy of Science. Vol. 104. Dordrecht,Boston,London. Kluwer Academic Publishers.

Hearst M. A. (1992) *Automatic Acquisition of Hyponyms from Large Text Corpora*. In: Proceedings of the 14th International Conference on Computational Linguistics COLING'92. pp 539-545.

Manning C. D., Schütze, H. (1999) *Foundations of Statistical Natural Language Processing*. Cambridge. MIT Press.

Mitchell T. M. (1997) *Machine Learning*, McGraw-Hill.

Muggleton S., De Raedt L. (1994) *Inductive Logic Programming: Theory and Methods*. In: Journal of Logic Programming. Vol.19-20. pp 629-679.

Oueslati R. (1999) *Aide à l'acquisition de connaissances à partir de corpus*. PhD thesis. Université Louis Pasteur. Strasbourg.

Pustejovsky J. (1995) *The Generative Lexicon*. Cambridge. MIT Press.