

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

Abstract

One relevant way to structure the 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. Methods from this former approach have been widely used and produce portable, robust, and fully automatic systems. They provide however poor explanations of their results, and may have difficulties to grasp very specific relations. The symbolic approach groups two strategies. The first one is the symbolic linguistic approach, in which operational definitions of the elements to acquire are manually established by linguists —usually in the form of morpho-lexical patterns that carry the studied terms or relations—, or by a list of linguistic clues. However, when such patterns or clues are unknown, but examples of elements respecting the target terms or relation are known, techniques from the second strategy of this symbolic approach can be used, *i.e.* symbolic machine learning (ML) methods. This facet of this approach, far less known and employed, is just beginning to appear and widen in the natural language processing community. The aim of this paper is to point out the interest of such techniques, and to show how they can be used to infer efficient and expressive extraction patterns of complex terms or lexical relations from examples of elements that verify the target relations or the form of the terms. However, these techniques are often supervised, *i.e.* require to be (manually) fed by examples. We also explain that one method from each of the numerical and symbolic ML approaches can be combined in order to keep advantages from both: meaningful patterns, efficient extraction and portability.

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

Abstract: One relevant way to structure the 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. Methods from this former approach have been widely used and produce portable, robust, and fully automatic systems. They provide however poor explanations of their results, and may have difficulties to grasp very specific relations. The symbolic approach groups two strategies. The first one is the symbolic linguistic approach, in which operational definitions of the elements to acquire are manually established by linguists —usually in the form of morpho-lexical patterns that carry the studied terms or relations—, or by a list of linguistic clues. However, when such patterns or clues are unknown, but examples of elements respecting the target terms or relation are known, techniques from the second strategy of this symbolic approach can be used, *i.e.* symbolic machine learning (ML) methods. This facet of this approach, far less known and employed, is just beginning to appear and widen in the natural language processing community. The aim of this paper is to point out the interest of such techniques, and to show how they can be used to infer efficient and expressive extraction patterns of complex terms or lexical relations from examples of elements that verify the target relations or the form of the terms. However, these techniques are often supervised, *i.e.* require to be (manually) fed by examples. We also explain that one method from each of the numerical and symbolic ML approaches can be combined in order to keep advantages from both: meaningful patterns, efficient extraction and portability.

1. Introduction

One relevant way to structure the domain of lexical knowledge 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. If the first kind of methods has been widely employed both to extract complex terms or syntagmatic, and paradigmatic relations [10], only one of the two strategies of the symbolic approach of acquisition has been really investigated: the symbolic linguistic approach, in which operational definitions of the elements to acquire are manually established by linguists —usually in the form of morpho-lexical patterns that carry the studied terms or relations—, or by a list of linguistic clues (*e.g.* [13]). However, when such patterns or clues are unknown¹, but examples of elements respecting the target terms or relation are known, symbolic machine learning (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 Hearst [9]:

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.*) [11] offers a framework to automate step 4, and automatically produce the unknown morpho-lexical patterns. In order to demonstrate what ML techniques can provide, this paper, through the description of an experiment concerning the acquisition of patterns of one type of semantic

¹ Or are domain-dependent.

relation, explains how one ML method, inductive logic programming (ILP, [12]) works, and what limits of numerical approaches such a technique can solve. After a first section dedicated to the key-points of numerical approaches, and their advantages and drawbacks, 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.

2. Numerical approach

Within the numerical approach, complex terms or relations between lexical units can be acquired by studying word cooccurrences in a text window (or a syntactic structure), and evaluating the strength of the association with the help of a statistical score (association coefficient) that detect words appearing together in a statistically significant way (*e.g.* [3]). Following Harris's linguistic principles [8], numerical distributional analysis methods respect a 3-step approach: extraction of the cooccurents 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 cooccurents (various measures are defined), clustering into classes, following different data analysis or graph techniques (*e.g.* [1, 7]).

Let us briefly sum up the advantages and drawbacks of numerical methods: they are portable 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.

3. ILP to acquire Noun-Verb relations

The use of symbolic ML methods is beginning to widen in the natural language processing community. Among these methods, ILP, thanks to its expressiveness and flexibility, has been applied to different problems (overview in [6]).

From examples and counter-examples of a concept, and a background knowledge, an ILP algorithm infers rules (Horn clauses) that cover (that is, characterize or explain) 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.

We have applied our ILP system [5], ASARES, to the acquisition of extraction patterns for some semantic relations between a noun(N) and a verb(V): the inferred rules or patterns must allowed us to extract N-V couples in which V plays the role of either the purpose or function of N (*e.g.* cut for knife), or its creation mode (build for house). Such N-V pairs are called *qualia-pairs* hereafter. Using a ML technique is especially well-suited here both because the extraction patterns are not known, and statistical cooccurrence-based methods have been proved not satisfactory for this task [2].

Here, the concept to be learned is the qualia nature of a N-V pair occurring within a sentence. An example (resp. counter-example) corresponds to a N-V couple manually indicated by an expert as verifying (resp. not verifying) the target relations in one sentence of a POS and semantically tagged corpus; the 3,000 examples and 3,000 counter-examples are represented by all the words and their tags occurring in their corresponding sentences. 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), artifact(N).

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, V is an action verb in the infinitive and N is an artifact.

As explained in [2], using the produced patterns to extract qualia N-V pairs from the corpus gives good results; moreover, the produced rules give access to a linguistically interpretable support to the target-concept.

Let us summarize the weak and strong points of symbolic approaches: they need *a priori* knowledge (e.g. examples), but produce interpretable results; detection is done at the occurrence level, and rare cases can be treated.

4. Concluding remarks

The cost of the ILP method, essentially lying in the construction by an expert of example and counter-example sets, makes it time-consuming, and thus difficult to apply to a new corpus. However, we have shown in [4] that it is possible to combine, in a so-called *semi-supervised* acquisition technique, one statistical cooccurrence-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—unsupervised aspect of the statistical acquisition, linguistically meaningful contextual pattern generation of the supervised symbolic one—and rival the performances of the former “pure” ILP system.

References

- [1] 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 IC'97, Ingénierie des Connaissances*, Roscoff, France, pp. 207-223.
- [2] Bouillon, P., Claveau, V., Fabre, C., and Sébillot, P., 2002, Acquisition of qualia elements from corpora - Evaluation of a symbolic learning method, *Proceedings of LREC'2002, 3rd International Conference on Language Resources and Evaluation*, Las Palmas, Canary Islands, Spain, pp. 208-215.
- [3] Church, K.W., and Hanks, P., 1989, Word association norms, mutual information, and lexicography, In: *Proceedings of ACL'89, 27th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada, pp. 76-83.
- [4] Claveau, V., and Sébillot, P., 2004, From efficiency to portability: acquisition of semantic relations by semi-supervised machine learning, In: *Proceedings of COLING'04, 20th International Conference on Computational Linguistics*, Geneva, Switzerland, pp. 261-267.
- [5] Claveau, V., Sébillot, P., Fabre, C., and Bouillon, P., 2003, Learning semantic lexicons from a part-of-speech and semantically tagged corpus using inductive logic programming, *Journal of Machine Learning Research, special issue on Inductive Logic Programming*, 4:493-525.
- [6] Cussens, J., and Džeroski, S., 2000, *Learning language in logic*, Vol. 1925, LNAI, Springer Verlag.
- [7] Grefenstette, G., 1994, *Explorations in Automatic Thesaurus Discovery*, Dordrecht: Kluwer Academic Publishers.
- [8] Harris, Z., Gottfried, M., Ryckman, T., Mattick, P.(Jr), Daladier, A., Harris, T.N., and Harris, S., 1989, *The Form of Information in Science, Analysis of Immunology Sublanguage*, Boston Studies in the Philosophy of Science, 104, Kluwer Academic Publisher, Dordrecht.
- [9] Hearst, M.A., 1992, Automatic acquisition of hyponyms from large text corpora, In: *Proceedings of COLING'92, 14th International Conference on Computational Linguistics*, Nantes, France, pp. 539-545.
- [10] Manning, C.D., and Schütze, H., 1999, *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, Massachusetts, USA.
- [11] Mitchell, T.M., 1997, *Machine Learning*, McGraw-Hill.
- [12] Muggleton, S., and De Raedt, L., 1994, Inductive logic programming: theory and methods, *Journal of Logic Programming*, 19-20:629-679.
- [13] Oueslati, R., 1999, *Aide à l'acquisition de connaissances à partir de corpus*, PhD thesis, Université Louis Pasteur, Strasbourg, France.