Knowledge acquisition of predicate argument structures from technical texts using Machine Learning: the system ASIUM

David Faure & Claire Nédellec

Laboratoire de Recherche en Informatique, UMR 86-23 du CNRS,
Équipe Inférence et Apprentissage,
Université Paris-Sud, bât 490, F-91405 Orsay,
{faure,cn}@lri.fr,
Tél {david, claire}: +33 (0)1.69.15.66.{07, 26}
Fax: +33 (0)1.69.15.65.86

Abstract. In this paper, we describe the Machine Learning system, ASIUM[1], which learns Subcaterorization Frames of verbs and ontologies from the syntactic parsing of technical texts in natural language. The restrictions of selection in the subcategorization frames are filled by the ontology's concepts. Applications requiring such knowledge are crucial and numerous. The most direct applications are semantic control of texts and syntactic parsing disambiguation.
This knowledge acquisition task cannot be fully automatically performed. Instead,we propose a cooperative ML method which provides the user with a global view of the acquisition task and also with acquisition tools like automatic concepts splitting, example generation, and an ontology view with attachments to the verbs. Validation steps using these features are intertwined with learning steps so that the user validates the concepts as they are learned. Experiments performed on two different corpora (cooking domain and patents) give very promising results.

Keywords: machine learning, natural language processing, ontology, predicate argument structure, corpus-based learning, clustering.

1 Introduction

Semantic knowledge acquisition from texts, such as predicate argument structures and ontologies is a crucial and difficult task and the manual acquisition is obviously long even in limited domains. New automatic methods involving both Natural Language Processing (NLP) and Machine Learning (ML) techniques ([Zelle93], among others) can give very good results in a short time. In this paper, we present ASIUM, a system that learns cooperatively from syntacticaly parsed texts without manual annotations, ontologies and subcategorization frames of verbs (SF) for specific domains following the principle of "domain dependence"[2] [Grefenstette92]. Subcategorization frames represent here a subcase of predicate argument structures where the predicate is restricted to a verb.

[1] Acquisition of SemantIc knowledge Using Machine learning methods.
[2] "*A semantic structure developed for one domain would not be applicable to another*".

ASIUM is based on an original unsupervised conceptual clustering method and, although the process cannot be fully automatized, provides interactive features in order to support the knowledge acquisition task.

We will show here how ASIUM is able to learn knowledge of good quality from possibly noisy texts and how ASIUM's cooperative features, together with its inductive capabilities, allow to acquire ontologies and *SF* in reasonable time[3].

2 Our approach

We attempt to acquire *SF* and ontologies from texts for texts control purposes for DASSAULT AVIATION company. Initially, we attempted to automatically revise and complete *SF* of a draft ontology manually acquired by a domain expert. This attemp failed for two main reasons: first the expert has too many a priori on the texts and second, he used incremental method to acquire the ontology. Revision of the acquired knowledge with respect to the training texts required profond reorganization of the ontology that incremental and even cooperative ML revision methods were not able to handle: it was locally consistent, but any revision leads to deeply restructuring it. This experiment illustrates one of the limitations of manual acquisition by domain experts without linguists and the need for knowledge acquisition tools.

Our aim is to *learn SF* and an ontology because no such bases were available. The few existing bases are too general and thus incomplete (EUROWORDNET or WORDNET). In a specific domain, the vocabulary as well as its possible usage are reduced, which makes such ontologies overly general. On the other hand, they may lack some specific terminology of the application domain.

As opposed to the approach consisting of completing and specializing general ontologies for specific domains as [Basili97] with WORDNET, the targeted approach we have chosen, even for English, is to learn suitable knowledge from a representative corpus of the domain, thus avoiding inconsistency risks.

3 Knowledge learned

ASIUM learns verb *SF* and *ontologies*. Here is an example of a *SF* for the verb `to inject`: `<to inject> <object: combustible> <in: furnace>`.

The two couples `<object: combustible>` and `<in: furnace>` are the *subcategories* of the verb `to inject`; `object` is a *syntactic role* and `in` is a preposition introducing an adjunct while `combustible` and `furnace` are their *restrictions of selection*. More generally a *SF* as ASIUM learns it, has the following form: `<verb> <syntactic role|preposition: concept*>*`.

The subcategories are arguments and adjuncts of the verb. In our framework, restrictions of selection (*RS*) can be filled with an exhaustive list of nouns (in canonical form) or by one or more *concepts* defined in

[3] About ten hours for the cooking domain of about 3 Mo of texts and 1120 verbs.

an ontology, where the meaning of the concepts is characterized by the *SF* they appear in. The ontology represents *generality relations* between concepts in the form of a directed acyclic graph. The axioms only express subsumption (IS-A) relationships between unary predicates or concepts. For instance, the ontology could define `fuel`, `gaz` and `carbon` as `combustible`, and `carbon` as both `combustible` and `burning wastes`. Our method learns such an ontology and *SF* in a cooperative and unsupervised (in the ML sense) manner from texts.

4 Overview of the method

The method implemented in the ASIUM system is included in a knowledge acquisition chain. It consists of the syntactic parser SYLEX [Constant95] providing ASIUM with all interpretations[4] of parsed sentences including attachments of noun phrases[5] to verbs and clauses, without any pre or postprocessing.

As a first step, ASIUM automatically extracts *instantiated subcategorization frames* from the syntactic parsing of clauses. The *instantiated SF* is similar to a *SF* but the *RS* are the actual head nouns occurring in the clause instead of concepts: `<verb> <prep. | syntactic role: head noun>*.`

Preliminary experiments show that instantiated *SF* are sufficient with respect to the learning task and that the ML method is robust with respect to parsing ambiguities or even failures.

The learning method relies on the observation of syntactic regularities in the context of words [Harris68]. We assume here that head nouns occuring with the same couple `verb+preposition/syntactic role` represent a so-called *basic class* and have a semantic similarity in the same line as [Grefenstette92], [Peat91] or others, but our method is based on a *double regularity model*: ASIUM gathers nouns together as representing a concept only if they share at least *two different* `(verb+preposition/syntactic role)` contexts as in [Grishman94]. Experiments show that it forms more reliable concepts, thus requiring less involvement from the user. Our similarity measure computes the overlap between two lists of nouns[6] (Details in [Faure98]). As usual in conceptual clustering, the validity of the concepts learned relies on the quality of the similarity measure between clusters which here increases with the size of their intersection.

Basic classes are then successively aggregated by a bottom-up breadth-first conceptual clustering method to form the concepts of the ontology level by level with expert validation and/or labelling at each level. Thus a given cluster cannot be used in a new construction before it has been validated. For complexity reasons, the number of clusters to be aggregated is restricted to two, but this does not affect the relevance of the learned concept as shown in [Faure98]. Verb *SF* are learned in

[4] In case of ambiguity, ASIUM takes all of them.

[5] Nouns phrases are reduced to head nouns (stopwords and adjectives are removed).

[6] $Sim(C_1, C_2) = 1$ for lists with the same nouns and $Sim(C_1, C_2) = 0$ for lists without any common nouns.

parallel so that each new concept fills the corresponding *RS* then resulting in the generalization of the initial synthetic frames which allows to cover examples which *did not occur* as such in texts. Thus, the clustering process does not only identify the lists of nouns occuring after the same `verb+preposition/function` but also augments this list by *induction*.

For example, from those instantiated *SF*, `<to travel> <subject: [father,neighbor,friend]><by:[car,train]>` and `<to drive> <subject: [friend,colleague]> <object: [car,motor-bike]>`, ASIUM learns both concepts `<Human>`, `<Motorized vehicle>` defined as `father,neighbor,friend,colleague` and `car,train,motor-bike` and both *SF*, `<to travel> <subject: Human> <by: Motorized vehicle>` and `<to drive> <subject: Human> <object: Motorized vehicle>`.

The risk of over-generalization is controlled both by a clustering threshold and the user. Concept learning could not be fully automated since the attachment of the concepts learned as *RS* of verbs must be validated by an expert in order to limit the risk of over-generality that the clustering threshold cannot completely avoid. Thus concept formation is intertwined with cooperative validation steps where the domain expert assesses and refines the learning results on line if needed, given acquisition tools like automatic concepts splitting, examples generation and ontology view with attachments to the verbs.

5 Experimentations

ASIUM has been applied first on a cooking recipe corpora in French with the aim of applying it to maintenance texts at DASSAULT AVIATION for language control purposes. Second, we have applied ASIUM on Oxy-fuel burner (a specific kind of burner using oxidants) patents for technical watch.

Evaluation of the unsupervised learned knowledge quality is a very difficult problem for which we have currently no solutions, but only highlights. First, ASIUM is included in a chain. Its efficiency could be partially measured by the utility and the improvement of the final task performance but once an error has been identified in the final output, locating the original faulty component is difficult in case an intermediate evaluation is not possible. Second, evaluating the cooperative system independently of the user is difficult. Third, the results of the learning process should be evaluated with respect to the quantity[7] and the nature of the user's work using counters on each type of action[8]. Counters will only give a partial view on the quality of the learned knowledge and the quality of the interaction tools and should be completed. Other evaluations of the quality of the results regarding redundancy of the corpora and of the induction effect in terms of completeness have been done in [Faure98]. They should be completed by correctness measures. As no negative example is available,

[7] Duration of the cooperative process regarding time needed in order to learn the same knowledge by hand.

[8] For instance, how many irrevelant inductions did the user refuse?

the measure of (verb+preposition/function+noun) induced from a training set and *not useful in a test set* could be a good indicator of correctness.

An evaluation of ASIUM results, done independently from a final application can not give a final answer to the evaluation question, only hits. For instance, the ontologies and *SF* learned could be compared to other lexicons but it would not only require the measurement of the similarity [Shaw89] but also the nature of the difference in case of a discrepancy.

6 Related Work

As proposed by [Hindle90] and [Pereira93], our method clusters nouns on the basis of syntactic regularities but without restricting the syntactic roles to be learned from subjects and objects. Our claim is that in technical domains the verbs are not only characterized by their arguments. Compared to [Grefenstette92], or [Bourigault96], ASIUM exploits two levels of regularities in the context instead of one. In ASIUM this would amount to learning basic classes as concepts which is obviously not suitable. [Brent91] learns the *SF* from large corpora from untagged texts with an automatic approach and focuses on learning five given *SF*. [Buchholz98] learns *SF* comparable to the ones learned by ASIUM with a supervised approach which is very time-consuming for the expert. In the same framework, WOLFIE [Thompson95] coupled with CHILL [Zelle93], learns case-roles and a lexicon from semantically annotated corpora by hand. Case-roles differ from *SF* as learned by ASIUM in that prepositions and syntactic roles are replaced by semantic roles such as agent or patient. Such information allows one to distinguish among the different semantic roles of given prepositions. As opposed to ASIUM ontology, the *RS* learned by WOLFIE are lists of attribute-values defining the concepts. Moreover WOLFIE requires that the input sentences parsed by CHILL are all annotated by semantic labels (roles and restrictions). Unsupervised learning, as in ASIUM, delays concept labeling after learning, thus reducing considerably the end-user task. In the same way, semantic roles could be labeled once ASIUM learns the *SF* by assuming that different restrictions (couples syntactic role/preposition+concept) reflect different semantic roles.

7 Conclusion

In this paper, we have presented a cooperative ML system, ASIUM, which is able to acquire subcategorization frames with restrictions of selection and ontology for specific domains from syntactically parsed technical texts in natural language. Texts and parsing may be noisy. The knowledge acquisition task is based on an original unsupervised clustering method. Needed expert validation and adjustment are supported by cooperative tools giving the expert a global and manageable view on the whole corpus helping him to integrate the needed domain knowledge that would not appear in the corpus.

Preliminary experiments on corpora of cooking recipes in French, and patents in English, have shown the applicability of the method to texts

in restricted and technical domains and the usefulness of the cooperative approach for such knowledge acquisition.

Further work will address evaluation aspects and semantic classes of verb learning from *SF* and ontologies.

Acknowledgement: This work is partially supported by the CEC through the ESPRIT contract LTR 20237 (ILP 2).

References

[Basili97] R. Basili and M. T. Pazienza. Lexical Acquisition for Information Extraction. In Maria Teresa Pazienza, editor, *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*, pages 14–18, Frascati, Italy, July 1997. LNAI Tutorial, Springer.

[Bourigault96] D. Bourigault, I. Gonzalez-Mullier, and C. Gros. LEXTER, a Natural Language Processing Tool for Terminology Extraction. In *7th EURALEX International Congress*, Göteborg, August 1996.

[Brent91] M. R. Brent. Automatic acquisition of subcategorization frames from untagged text. In *Proceedings of the 29st annual meeting of the Association for Computational Linguistics, ACL*, pages 209–214, 1991.

[Buchholz98] S. Buchholz. Distinguishing Complements from Adjuncts using Memory-Based Learning. In *Proceedings of the ESSLLI'98 workshop on Automated Acquisition of Syntax and Parsing*, 1998.

[Constant95] P. Constant. L'analyseur Linguistique SYLEX. In *5ème École d'été du CNET*, 1995.

[Faure98] D. Faure and C. Nédellec. A Corpus-based Conceptual Clustering Method for Verb Frames and Ontology Acquisition. In Paola Velardi, editor, *LREC workshop on Adapting lexical and corpus ressources to sublanguages and applications*, pages 5–12, Granada, Spain, May 1998.

[Grefenstette92] G. Grefenstette. Sextant: exploring unexplored contexts for semantic extraction from syntactic analysis. In *Proceedings of the 30st annual meeting of the Association for Computational Linguistics, ACL*, 1992. 14-18.

[Grishman94] R. Grishman and J. Sterling. Generalizing Automatically Generated Selectional Patterns. *Proceedings of COLING '94 15th International Conference on Computational Linguistics, Kyoto, Japan*, August 1994.

[Harris68] Z. Harris. *Mathematical Structures of Language*. New York: Wiley, 1968.

[Hindle90] D. Hindle. Noun classification from predicate-argument structures. In *Proceedings of the 28st annual meeting of the Association for Computational Linguistics, ACL, Pittsburgh, PA*, pages 1268–1275, 1990.

[Peat91] H.J. Peat and P. Willet. The limitations of term co-occurrence data for query expansion in document retrieval systems. *Journal of the American Society for Information Science*, 42(5):378–383, 1991.

[Pereira93] F. Pereira, N. Tishby, and L. Lee. Distributional Clustering of English Words. In *Proceedings of the 31st annual meeting of the Association for Computational Linguistics, ACL*, pages 183–190, 1993.

[Shaw89] M.L.G. Shaw and B. R. Gaines. Comparing conceptual structures: consensus, conflict, correspondence and contrast. In *Knowledge Acquisition*, volume 1, pages 341–363, 1989.

[Thompson95] C. A. Thompson. Acquisition of a Lexicon from Semantic Representations of Sentences. In *33rd Annual Meeting of the Association of Computational Linguistics, Boston, MA July, (ACL-95).*, pages 335–337, 1995.

[Zelle93] J. M. Zelle and R. J. Mooney. Learning semantic grammars with constructive inductive logic programming. *Proceedings of the Eleventh National Conference on Artificial Intelligence*, pages 817–822, 1993.