ASIUM: learning subcategorization frames and restrictions of selection

David Faure[†*] & Claire Nédellec[*]

[*]Laboratoire de Recherche en Informatique, URA 410 du CNRS, Équipe Inférence et Apprentissage, Université Paris Sud, bât 490 F-91405 Orsay {faure,cn}@lri.fr

[†]Laboratoire d'Intelligence Artificielle de Paris V, LIAP-5, Université René Descartes, Centre universitaire des Saints pères, 45 rue des Saints pères, F-75006 Paris faure@descartes.math-info.univ-paris5.fr

Abstract

We describe in this paper the ML system, ASIUM, which learns subcategorization frames of verbs and ontologies from syntactic parsing of technical texts in natural language. The restrictions of selection in the subcategorization frames are filled by the concepts of the ontology. Applications requiring subcategorization frames and ontologies are crucial and numerous. The most direct applications are semantic checking of texts and syntactic parsing improvement but also text generation and translation. The input of ASIUM result from syntactic parsing of texts, they are subcategorization examples and basic clusters formed by head words that occur with the same verb after the same preposition (or with the same syntactical role). ASIUM successively aggregates the clusters to form new concepts in the form of a generality graph that represents the ontology of the domain. Subcategorization frames are learned in parallel, so that as concepts are formed, they fill restrictions of selection in the subcategorization frames. Asium method is based on conceptual clustering. First experiments have been performed on a corpus of cooking recipes and give very promising results reported here.

Keywords: Natural language processing, semantic knowledge acquisition, clustering.

1 Introduction

The acquisition of semantic knowledge from texts is a crucial and difficult task. Increasing interest in the acquisition of semantic knowledge from large textual datasets, (or corpus), leads to the development of new automatic methods involving both NLP and ML techniques. In this article, we will present the system ASIUM, that, coupled with a syntactic parser learns subcategorization frames and ontologies from texts in natural language without requiring any annotation of texts by hand. ASIUM is based on an original unsupervised conceptual clustering method. It has been applied on a cooking recipe corpus in French with the view to apply it then to maintenance texts of DASSAULT AVIATION company. We will show here how ASIUM is able to learn knowledge of high quality from possibly noisy texts and thus to save considerable amount of time compared to acquisition by hand.

1.1 Semantic knowledge learned by ASIUM

ASIUM learns semantic knowledge in the form of *subcategorization frames* of verbs and *ontologies*. Here is an example of subcategorization frame for the verb `to travel`: `<to travel> <subject: human> <by: vehicle>`. The two couples `<subject: human>` and `<by: vehicle>` are the subcategories of the verb `to travel`. `subject` is a syntactic role and `by` is a preposition introducing an adjunct while `human` and `vehicle` are their *restrictions of selection*. More generally a subcategorization frame as Asium learns it, has the following form, `<verb> <syntactic role|preposition: noun|concept>*`.

The subcategories are arguments of the verb (subject, direct object or indirect object) and adjuncts. In our framework, restrictions of selection can be filled by an exhaustive list of nouns (in canonical form) or by one or more concepts defined in an ontology. The ontology represents *generality relations* between concepts in the form of an acyclic oriented graph. For instance, the ontology could define car, train and motorcycle as motorized vehicle, and motorized vehicle as both vehicle and pollutant. Our method learns such an ontology and subcategorization frames in an unsupervised manner from texts in natural language. The concepts formed have to be labeled by an expert.

1.2 Potential applications

Applications of such semantic knowledge are numerous in text understanding. Let us give some examples. Subcategorization frames and ontologies can be learned from corpora and then new texts can be semantically parsed by automatically mapping subcategorization frames to clauses that have been syntactically parsed. For example, the clause My father travels by car. maps to <to travel> <subject: father, man, human> <by: car, motorized vehicle, vehicle>. This type of semantic interpretation is a first step towards building semantic models of texts. Such a semantic parsing can also improves syntactic parsing by removing ambiguities like in Cooking 3 minutes where the subcategorization frame of to cook will lead to interpret 3 minutes as an adverb instead of as a direct object. Semantic parsing has also direct applications in *text categorization*, *translation*, *text generation* or can be used for *MRD filling*.

Manual acquisition of such semantic knowledge is obviously long and difficult even in limited domains. However, acquisition from texts by machine learning methods can give very good results in short time without requiring tedious annotations. We have developed and applied such methods on technical texts in collaboration with DASSAULT AVIATION company. Their texts have the following properties compared to texts in general domain: they are domain specific, the vocabulary is thus limited, the polysemia is restricted and verbs are mostly concrete and action verbs.

Reducing the ambiguities or detecting semantic errors in maintenance texts is one of the main goal of this application. Subcategorization frames and ontologies learned from training texts would fill the knowledge base of a writer-assistant which would then prompt the expert when incorrect or ambiguous clauses are detected in unseen maintenance texts. Incorrect or ambiguous clauses are clauses that do not map the subcategorization frame of their verb. For example the subcategorization frame: <to call> <subject: human> <object: human> allows to detect a forbidden metonymy: The control tower calls the pilot (We suppose that Control tower is defined as Building in the ontology. Depending on the required precision degree, the subject of to call will be restricted to human or not.

The kind of ontology we need is not as general as defined in the AI literature: "specification of a conceptualization" [Gru95] (in other words, an ontology is an axiomatic characterization of the meaning of a logical vocabulary). In many cases such as ours, the axioms of the ontology only express subsumption (IS-A) relationships between unary predicates or concepts. However, you may notice that the meaning of concepts of the ontology is characterized here by the subcategorization frames they appear in.

Section 2 presents an overview of the methods applied to acquire subcategorization frames and ontology from texts. Section 3 details the clustering techniques it is based on. Section 4 describes the cooperative features which are required to label the clusters as concepts and to palliate syntactic parsing errors and the lack of representativeness of texts. Section 5 presents preliminary experimental results.

2 Overview of the methods

2.1 Available subcategorization frames and ontologies

No subcategorization frame base nor ontology were available in French which could be useful for our application with respect to the domain. The few existing bases are too general and thus incomplete. For example, WordNet[Milty], a very large ontology of about 60.000 English words, is not yet available for French (the project "EuroWordNet" [Eur96] is adapting WordNet to other European languages). Automatic translation of subparts of WordNet by using Machine Readable Dictionary (MRD) ([Bri88], [Wil97]) may at first sight be a reasonable solution, but it raises numerous problems. Among others, some English words may have more than one translation in French, while some other may have no direct translation.

But a more serious objection raises against the use of a general ontology such as WordNet. This ontology, although very complete, is not suitable for processing text in technical language. On the one hand WordNet is not a purpose directed ontology, it may store up to seven meanings and syntactic roles for a word increasing thus the risk of semantic ambiguity. In a specific domain, the vocabulary as well as its possible usage are reduced, which makes an ontology such as WordNet overly general. On the other hand, WordNet may lack some specific terminology of the application domain.

Attempts to automatically revise subcategorization frames and a subset of an ontology acquired by a domain expert of Dassault Aviation have failed. Revision of the acquired knowledge with respect to the training texts required deep restructuration of the knowledge that incremental and even cooperative ML revision methods where not able to handle. The main reason is that the expert has built the ontology and the subcategorization frames with too many a priori that were not reflected in the texts. This experiment illustrates one of the limitation of manual acquisition by domain experts without linguists.

2.2 Learning subcategorization frames and ontologies

We have thus designed and developed a complete knowledge acquisition chain including unsupervised cooperative machine learning methods which takes training texts in natural language as input and built the desired subcategorization frame base and ontology. Thus, acquired knowledge exactly reflects the linguistic phenomena occurring in the texts. Sensibility to the quality of training texts is the counterpart. The acquisition chain consists of the syntactic parser Sylex [Con95] coupled to Asium. Sylex provides Asium with parsed sentences including attachments of noun phrases to verbs and clauses. As a first step, Asium automatically extracts *instantiated subcategorization frames* from syntactic parsing of clauses. The *instantiated subcategorization frame* of a clause is similar to a subcategorization frames but the restrictions of selection are the actual head words (in canonical form) occurring in the clause instead of concepts.

`<verb> <preposition | syntactical role: head word>*`

For example, the instantiated subcategorization frame of the clause `My father travels by car` is: `<to travel> <subject: father> <by: car>`. Notice that stopwords [Tes88] and adjectives are removed, only head words (main nouns in arguments and adjuncts) appear in instantiated subcategorization frames. Preliminary experiments show that this information is sufficient with respect to the semantic task. Moreover the lexicon of the syntactic parser identifies if the head words are expressions instead of single words.

In case of syntaxic ambiguities, Sylex gives all the differents interpretations and ASIUM uses all theses interpretations. Experiments have shown that the machine learning method works well with theses ambiguities and acquisition of semantic knowledge is not affected. This method avoids a very time-comsuming hand desambiguation step.

The learning method takes the instanciated frames as input and learns the ontology and the subcategorization frame base. The method is based on an unsupervised clustering method and relies on the following assumption: *head words occurring after the same,* *different prepositions (or with the same, different syntactic roles), and with the same, different verbs represent a same concept.* For instance, let us suppose the nouns car, train and motorcycle occur in different clauses as adjunct of the verb to travel, after the preposition by, and also as direct object of the verb to drive, these nouns are thus considered as representing a same concept. In other words, the more often they occur in the same context, the more reliable is the underlying concept. The context here is the only syntactical roles, or prepositions, of the head words and the verbs they are attached to. Relying on this assumption, the learning process successively aggregates such sets of words into clusters at different generality levels to form concepts of the ontology and fill the restrictions of the verb. This assumption seems to be more reliable in technical texts where the restrictions of selection of verbs are more restricted and where general verbs such as "to present" or "to mean" are less numerous than specific verbs like "to ignit" or "to land".

This assumption is concretized in the learning method which consists in two steps, factorization and clustering. The first step consists in gathering head words that occur in the same contexts, i.e. with the same verb and the same preposition / syntactic role. These sets of words form the so-called *basic classes.* This comes to factorize the instantiated subcategorization frames into so-called *synthetic* frames according to their verbs. Its number of occurrences in the context is associated to each head word. For example, from those instantiated subcategorization frames,

<to travel> <subject: father> <by: car>
<to travel> <subject: neighbor> <by: train>
<to drive> <subject: friend> <object: car>
<to drive> <subject: colleague> <object: motorbike>
<to drive> <subject: friend> <object: motorbike>

ASIUM creates two synthetic frames, one per verb:

<to travel> <subject: [father(1), neighbor(1)]> <by: [car(1), train(1))]> and <to drive> <subject: [friend(2), colleague(1)]> <object: [car(1), motorbike(2)]

Basics classes are then successively aggregated by the conceptual clustering method to form the concepts of the ontology (§ 3). As usual in conceptual clustering, the validity of the new concepts relies on the quality of the similarity measure between clusters which increases with the size of their intersection. Subcategorization frames of verbs are learned as concepts are formed so that each new concept fills the corresponding restriction of selection of the frame resulting then in the generalization of the synthetic frames. Subcategorization frames learned with the previous examples could be:

<to travel> <subject: human>) <by: motorized vehicle>
<to drive> <subject: human>) <object: motorized vehicle>

The formation of the new concept motorized vehicle from the two basic classes, [car(1), train(1))] and [car(1), motorbike(2)] yields the generalization of the synthetic frames of to travel and to drive. The frame of to travel now admits motorbike as adjunct after the by preposition and the frame of to drive now admits train as object, examples which did not occurs as such in texts.

Clustering steps are intertwined with cooperative validation steps where a domain expert assesses and refines the learning results (§ 4) on line if needed, given the graphic and user-friendly interface of ASIUM.

3 Conceptual clustering method

3.1 Existing methods

Incremental ascending existing methods such as COBWEB [Fis87] which take vectors as input examples are not suitable for complexity reasons. In our case, examples to cluster are sets of words, associated to the frequency of the corresponding instantiated frame in the corpus. Attributes of the input vector would be head words and values, their frequencies.

As attributes would need to be the same in all vectors, very large vectors representing a whole dictionary would be required (about 2000 words in our experimentation) and most of their values would be equal to zero. AUTOCLASS [Che88], CLASSIT [Gen88], or ADECLU [Dec91] have the same drawback. FOL-based clustering method such as KBG [Bis92] or RIBL [Emd96] use a more powerful representation than we need but those methods have the following limitations for our approach. They learn strict hierarchies of concepts although to express different viewpoints in an ontology a semantic class may have more than one super-class. Moreover their last clustering steps create over general clusters (the root concepts) which are useless as far as semantic checking is concerned.

3.2 Description of Asium clustering method

We have thus developed a conceptual clustering method suitable for learning ontologies and subcategorization frames of verbs. ASIUM (see algorithm fig. 1) is bottom-up and performs best-first. It takes basics classes as input as defined above and builds the ontology level by level. For complexity reasons, the number of clusters to be aggregated is restricted to two, but it does not affect the relevance of the learned concept as shown below. As hierarchy is too restricted as representation to express the complexity of the ontologies in the domains we have studied, ASIUM builds acyclic oriented graph where links between concepts represent the generality relation.

The first clustering step builds the first level of the ontology. Distances between all pairs of clusters are computed. Two basic classes are aggregated if the distance is less than threshold set by the user. The distance is defined in paragraph 3.3). The same way, a step n builds the level n of the ontology by aggregating pairs of clusters the distances of which are less than the threshold *without* taking into account their level. Clusters of level n can be aggregated to other cluster when level n is finished only. An additional but obvious constraint avoids the aggregation of a cluster with one of its descendant in the generality graph. This constraint ensures algorithm termination. The distinction

$Clusters\_to\_Aggregate \leftarrow Basic\_Classes.$
$New\_Clusters \leftarrow Basic\_Classes.$
$level \leftarrow 1.$
Repeat
    $Candidate\_Clusters \leftarrow \emptyset.$
    for all clusters (C_i, C_j), $C_i \in New\_Clusters$ and
    $C_j \in Clusters\_to\_Aggregate$ which verify the non-descendance constraint
        if $dist(C_i, C_j) < Threshold$ **and** $C_i \neq C_j$
            then
            $C_{new} \leftarrow aggregate(C_i, C_j)$
            $Candidate\_Clusters \leftarrow Candidate\_Clusters \cup C_{new}$
    endfor
    $New\_Clusters \leftarrow$ Cooperative validation of $Candidate\_Clusters$
    $Clusters\_to\_Aggregate \leftarrow Clusters\_to\_Aggregate \cup New\_Clusters.$
    $level \leftarrow level + 1.$
until $New\_Clusters = \emptyset$

Figure 1: Clustering algorithm

between $Clusters\_to\_Aggregate$ and $New\_Clusters$ avoids to recompute distances. After all admissible aggregations have been performed to create a new level of the ontology, the user validates all the learned clusters at that level. The lists of $Clusters\_to\_Aggregate$ and $New\_Clusters$ are updated and the process is repeated until no cluster can be aggregated anymore.

Restricting the number of clusters to be aggregated to pairs may lead to generate unuseful clusters from a conceptual point of view. In fact a post-processing tool removes all unuseful clusters, i.e. cluster not appearing as restrictions of selection in subcategorization frames.

Some comparisons have been done between the breadth-first strategy described here and a best- first method such that new clusters can be reused as soon as formed. The breadth-first strategy learns a more complete ontology: it tends to create smaller clusters than best-first and it thus preferable.

3.3 Definition of the distance

The relevance of the concepts formed by the clustering method strongly relies on the definition of an appropriate distance *dist* which compute the similarity between clusters [Rad89] and [Liu96]. *Dist* determines which pairs of clusters have to be aggregated into a new cluster which will replace them in all subcategorization frames where they occurred. Following our assumption, clusters which have a maximum overlap have to be merged in order to form new classes. Thus, clusters which contain the same words with the same frequencies are strictly similar (their distance is equal to 0), while the distance between disjoint clusters without any word in common is maximum, that is to say, equal to 1. *Dist* is defined as the proportion of common head words in the two clusters taking into account their frequency. The distance between cluster C_1 and C_2 is defined by:

$$dist(C_1, C_2) = 1 - [\frac{\sum FC_1 * \frac{Ncomm}{card(C_1)} + \sum FC_2 * \frac{Ncomm}{card(C_2)}}{\sum_{i=1}^{card(C_1)} f(word_{iC_1}) + \sum_{i=1}^{card(C_2)} f(word_{iC_2})}]$$

where $card(C_1)$ and $card(C_2)$ represent the number of different head words in clusters C_1 and C_2, and $Ncomm$, the number of different head words common to both C_1 and C_2. $\sum FC_1$ (resp. $\sum FC_2$) is the sum of the frequencies of the head words of C_1 (resp. C_2) also occurring in C_2 (resp.) C_1. $word_{iC_1}$ (resp. $word_{iC_2}$) is the i-th head word of cluster C_1 (resp. C_2), and $f(word_{iC_1})$ (resp. $f(word_{iC_2})$) is its frequency. The weights $\frac{Ncomm}{card(C_1)}$ and $\frac{Ncomm}{card(C_2)}$ minimize the influence of word frequencies, by offsetting the attraction phenomenon of very frequent words in subcategorization frames and increase the clustering efficiency. For example, two clusters:

C_1 : *to cook in*	C_2 : *to put in*

`oven(4)`
`stew pan(12)`
`frying pan(2)`
	`oven(5)`
	`stew pan(3)`
	`wok(6)`
	`pan(2)`

$\frac{Ncomm}{card(C_1)} = \frac{2}{3}$ (2 of the 3 head words of C_1 also occur in C_2 (`oven` and `stew pan`). $\frac{Ncomm}{card(C_2)} = \frac{2}{4}$. $\sum FC_1 = 4 + 12 = 16$, and $\sum FC_2 = 8$. thus, the distance is equal to:

$$1 - [\frac{(16 * \frac{2}{3}) + (8 * \frac{2}{4})}{(4 + 12 + 2) + (5 + 3 + 6 + 2)}] = 57\%$$

Dist is metric (the triangular inequality is still to be proved). It is inspired by the Hamming distance, and the "O-SIM" distance of KBG [Bis92], except that the frequency of words in contexts is taken into account. *Dist* has been adjusted with the recipe corpus on which it gives promising results (section 5).

3.4 Subcategorization frames learning

In parallel with concept formation, subcategorization frames are generalized so that new concepts replace their descendants in the restrictions of selection. In other words, basic classes in initial synthetic frames are successively replaced by the aggregated clusters as they are built. Each time a new cluster C is formed from two clusters C_1 and C_2, then C_1 and C_2 are generalized into C in all subcategorization frames where C_1 et C_2 occur as restrictions of selection. This way, in the travel example, the aggregation of clusters `[car(1), train(1))]` et `[car(1), motorbike(2)]` into a new cluster called `motorized vehicle` leads to the generalization of `to travel` and `to drive` frames. Aggregation can be fully automated while attachment of a new concept as the restriction of a verb must be validated by an expert. For instance, let us suppose that `bike(1)` belongs to the basic class of `to travel`, automatic aggregation will yield the new cluster `[car(1), train(1),` *bike(1)* `motorbike(2)]`. This generalization is relevant to the verb `to travel` (`motorbike(2)` is added), but irrelevant for the verb `to drive`, which requires a motorized vehicle. General concepts may be relevant for given verb while they are too general for others although they are built from.

4 Cooperative Learning

The participation of the user is needed in such a method, not only to control the generality level of restrictions in verb frames but also to interactively correct the clusters in case of noise. Such a clustering method applied to real texts is sensitive to the quality of the training set which is never fully correct in real applications. We have developed a cooperative interface which allows the user to both control the learning process and provide ASIUM with new domain knowledge. The interface provides inspection and refinement features which gives the user a manageable view of the knowledge base so that he can take the appropriate decisions.

The main points regarding user cooperation are the following. By labeling himself the clusters as they are built, the user will get a more comprehensible ontology. In [Tou97], the system automatically labels the learned clusters, by using the most frequent word in the cluster. This approach does not seem appropriate for our problem. Among other reasons, several clusters may have the same most frequent word (especially as our clusters may be overlapping), or there may be clusters in which several words have the highest frequency. However, in text control or syntactical parsing improvement, concept labeling is not needed as the definition of concept in extension (i.e. by listing head words) is sufficient.

The expert user validates the new clusters level per level. Thus, a given cluster cannot be used as subset of a new cluster before validation. Intertwining validation and learning in such a way guarantees the relevancy of learned concept while post validation would require deep revisions to be done by hand. The clusters are displayed in similarity order so that it may be easier to set the threshold. The validation step (acceptation / rejection) of clusters can be completed by adjustment operations. The main ones are rejection of given words as restrictions of selection of given verbs, and partition of new clusters into subclusters that would not have been identified before that point (see section 4.2). Validation and adjustment by the expert are difficult knowledge acquisition tasks the quality of the results depend on.

In order to support this task, we have developed cooperative tools inspired by APT [Ned96a], HAIKU [Ned96b] and EDINOS / REVINOS [Poi97] methods and specially a method to generate examples inspired by the system APT. Each tool benefits from the user friendly interface needed when used by non computer scientists.

4.1 Subcategorization frame adjustment

It is easier sometimes for an expert to classify "examples of what is learned" instead of evaluating and adjusting general frames or deciding if a given cluster can fill the restriction of selection of a verb frame. Examples here are "clauses" in the sense of instanciations of frames to be validated where restrictions are words of the new cluster, such as for instance, `to drive car` for validating the attachment of `car` as `motorized vehicle` to the verb `to drive`. One of the ASIUM's cooperative tools automatically generate such examples on demand in a similar way as APT system [Ned96a].

Following the ideas illustrated in APT, it is easier for an expert to classify an example of a concept (here, an instantiated verb frame) than to assess a verb frame, or to evaluate if a given cluster may fill a verb semantic feature and generalize it. Thus, when the user has to validate new clusters as acceptable or over-general as semantic features of given verbs, ASIUM leaves him the possibility to display examples of "sentences" which would be covered by the generalized syntactic frames. In fact, theses sentences are instanciations of the frame to be validated with as semantic features words belonging to the new learned cluster.

For validating the use of words of a new cluster in a given verb frame, the user has the possibility of asking for *newly* covered examples or for *all* covered examples, including examples found in the corpus. Newly covered examples are clause examples with words that belong to the new cluster without belonging to its descendant clusters. For instance, the newly covered examples for validating the new concept `vehicle` and its attachment as restriction of selection of the verbs `to travel` and `to drive` will be: `to travel by motorbike`, `to drive train` and `to drive bike`. By classifying the examples as positive and negative the user adjusts the automatic aggregation so that some words are removed from the generalization for some verbs. For instance `to drive bike` should rejected and as a consequence, `bike` is removed from the cluster attached to `to drive`. In that case, the new cluster is attached as such `to travel` and another one is formed for `to drive` without `to bike`. By inspecting `all examples`, the user has the opportunity to detect errors in the corpus or parsing errors. To refuse an example issued from the corpus is useful if it is an error of the syntaxic parser or if it is an use of a term the expert wants to forbid.

Other cooperative tools support cluster partition and labeling, word re-spelling, direct removing of irrelevant words in new clusters, and propagation to descendant clusters. Inspection facilities display the ontology and the concepts attachments to verbs.

4.2 Partition of basic classes

We have assumed until here that the generation of clusters leads to the identification of all relevant concepts. In fact it may be the case that clusters contain sub-concepts that the corpus did not enable to exhibit. The partition option of ASIUM automatically partitions basic classes before the clustering phase so that intersecting clusters are partitioned into three new clusters, the two complementary sets and the intersection, (in case of inclusion, the complementary set only is extracted). Two parameters are available to restrict the number of new clusters: the *minimal cardinal for the intersection of the two overlapping clusters* and the *minimum cardinal of the new cluster*. The user may also choose to remove or to keep the source clusters. In any case, at any time during the clustering phase, the user may partition the candidate cluster and label the sub-clusters.

5 Experimentation

As preliminary experimentation, we have applied ASIUM to cooking recipes in French because maintenance texts are not available outside DASSAULT AVIATION site. Recipes present the same linguistic phenomena as DASSAULT AVIATION's texts with respect to the semantic acquisition task and represent thus a good basis for first experiments. Moreover, one advantage of this application domain is that we are (relative) experts of this domain, and can take part in the validation process of the knowledge learned by ASIUM. Our experiments applied on a 3 Mo. corpus[1] gathered from the web. It contains around 1500 recipes. More than 1000 verbs occur in 90 000 clauses.

[1] http://www.lri.fr/Francais/Recherche/ia/sujets/langnat.html

SYLEX has been applied to the whole corpus (frontiers between recipes are not relevant here). From syntactic parsing ASIUM builds around 2300 basic clusters and 1000 synthetic verb frames (one per verb). With a similarity threshold of 40 % and if all created clusters are validated as relevant by the user, ASIUM generates only 239 clusters spread over the ontology levels as reported in the left table:

ontology level	number of clusters
1	44
2	130
3	58
4	17

Training set, (%))	% accuracy
90	99.53
70	97.10
50	92.10
30	82.57
10	60.63
5	47.87
1	26.87

The quality of what is learned is very high: most of the clusters are relevant and require few adjustments before filling restrictions of verbs. Most of the needed adjustments are due to corpus or parsing errors. The most frequent user action is cluster partition. For instance, among 239 concepts, ASIUM learns the concept of `liquid`, of `container` and of `physical measure` (`heat`, `pressure`, etc.). The acquisition of the ontology and verb frames takes only few hours as the number of concepts is low (249 for 1000 verbs). These good results have to be confirmed with other applications and the limitations of the method have to be precised. In particular, we have to evaluate the sensitivity of the method to the type of texts (more or less centered on a domain, where the verb are more or less specific) and to parsing errors which could be more numerous in more complex texts.

The size of the corpus needed to learn such knowledge obviously depends on the quality required by the application. The representativeness of the corpus has been evaluated with respect to the method. ASIUM has learned basic classes from a training subset of the corpus and then we have computed the proportion of couples verb-argument/-adjunct covered by the learned knowledge in the rest of the corpus. The results are reported above in the right table. 3 Mo. corpus appears as highly redundant with respect to the learning task, since only 5 % (150 Ko.) of the corpus is sufficient to cover 48 % in the remaining 95 % of the corpus. Our intuition is that using the adjuncts of the clause and not only of the arguments of the verb (subject, direct object and indirect object) highlights the regularities among sets of head words. As opposed to one could expect, adjuncts in cooking recipes are as constrained as arguments. This explanation has to be confirmed by other corpora.

6 Related Work

As proposed by [Hin90] and [Per93], our method clusters nouns on the basis of syntactic regularities observed in a corpus, but they restrict the syntactic roles to learn from, to subjects and objects of the verb. Our claim is that in technical domains the verbs are characterized by all their complements and not only by their arguments. With this additional input, smaller corpora are thus needed to observe the desired regularities.

WOLFIE [Tho95] coupled to CHILL [Zel93], learns case-roles and a lexicon from semantically annotated corpora. Case-roles differs from subcategorization frames as learned by ASIUM in that prepositions and syntactic roles are replaced by semantic roles such as `agent` or `patient`. Such information allows to distinghuish between the different semantic roles of given prepositions. As opposed to ASIUM ontology, the restrictions of selection 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, reducing thus considerably the end-user task. The same way, semantic roles could be labeled once the subcategorization frames learned by ASIUM by assuming that different restrictions reflect different semantic roles.

7 Conclusion

We have presented in this paper ASIUM, a cooperative Machine Learning system which is able to acquire subcategorization frames with restrictions of selection and ontologies for specific domains from syntactically parsed technical texts in natural language. Texts and parsing may be noisy. It is based on an original unsupervised breadth-first clustering method suitable for such an acquisition. Needed knowledge validation and adjustment is supported by cooperative tools such as automatic example generation.

Preliminary experiments with a corpus of cooking recipes has shown the applicability of the method to texts in restricted and technical domains. Further work will be done to validate the approach. In particular application to DASSAULT AVIATION maintenance texts will allow to evaluate the cooperative aspects of the system with experts who are not familiar with Machine Learning.

As a first step, ASIUM is applied to maintenance texts in order to fill dictionaries (MRD) for improving a syntactic parser. Semantic classes of verbs should be then learnable from verb frames and ontologies by applying FOL clustering methods such as KBG [Bis92]), by assuming that two verbs with the same frame must belong to the same semantic class, thus following [Bas96] assumption.

Acknowledgments

This work is partially supported by the CEC through the ESPRIT contract LTR 20237 (ILP 2). The authors wish to thank Celine Rouveirol of our group, Olivier Sigaud and Farid Cerbah of DASSAULT AVIATION Company for their active collaboration on this work.

References

[Bas96] Basili R., Pazienza M. T., Velardi P. *A Context Driven Conceptual Clustering Method for Verb Classification*, volume Corpus Processing for Lexical Acquisition. The MIT Press, Cambridge, Massachusetts, Branimir Boguraev, James Pustejovsky edition, 1996.

[Bis92] Bisson G. Learning in fol with a similarity measure. In *Tenth National Conference on Artificial Intelligence*, San Jose, California, 1992.

[Bri88] Briscoe E. J. and Boguraev B. K. *Computational Lexicography for Natural Language Processing*. Longman/Wiley, London/New York, 1988.

[Che88] Cheeseman P. and Kelly J. and Self M. and Stutz J. and Taylor W. and Freeman D. AutoClass: a Bayesian classification system. In *Proceedings of IWML'88, Ann Arbor*, pages 54–64, San Mateo, CA, 1988. Morgan Kaufmann.

[Con95] Constant P. L'analyseur linguistique SYLEX. In 5^{me} *École d'été du CNET*, 1995.

[Dec91] Decaestecker C. Description Contrasting in Incremental Concept Formation. In Y. Kodratoff, editor, *Proceedings of the European Working Session on Learning : Machine Learning (EWSL-91)*, volume 482 of *LNAI*, pages 220–233, Porto, Portugal, March 1991. Springer Verlag.

[Emd96] Emde W. and Wettschereck D. Relational Instance Based Learning. In Lorenza Saitta, editor, *ICML'96*, pages 122–130. Morgan Kaufmann Publishers, 1996.

[Eur96] Eurowordnet:building a multilingual database with wordnets for several european languages. http://www.let.uva.nl/~ewn/, March 1996.

[Fis87] Fisher D. H. Knowledge Acquisition Via Incremental Conceptual Clustering. *Machine Learning*, 2(2):139,172, September 1987.

[Gen88] Gennari J. H. and Langley P. and Fisher D. Models of Incremental Concept Formation. Technical Report ICS-TR-88-16, University of California, Irvine, Department of Information and Computer Science, June 1988.

[Gru95] Gruber, T. Towards Principles for the Design of Ontologies Used for Knowledge Sharing. In *Knowledge Acquisition*, pages 43(5/6): 907–928. International Journal of Human and Computer Studies, 1995.

[Hin90] Hindle D. 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.

[Liu96] Liu W. Z. An Integrated Approach for Different Attribute Types in Nearest Neighbour Classification. *The Knowledge Engineering Review*, 1996.

[Milty] Miller et al. Wordnet 1.5. Available on ftp.ims.uni-stuttgart.de /pub/WordNet/, 1990-1993, Princeton University.

[Ned96a] Nedellec C. APT a cooperative ML System. In A. Cypher, Y. Gil, and Pazzani M., editors, *Proceedings of AAAI Spring Symposium on Learning by demonstration.* AAAI Press, March 1996.

[Ned96b] Nedellec C., Rouveirol C., Ade H., Bergadano F., and Tausend B. *Advances in Inductive Logic Programming*, chapter Declarative Bias in Inductive Logic Programming, pages 82–103. Raedt de L. (ed.), IOS Press, 1996.

[Per93] Pereira, F., 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.

[Poi97] Poittevin L. Revinos: An interactive revision tool based on the concept of situation. In E. Plaza and Benjamins R, editors, *10th European Knowledge Acquisition Modeling and Management Workshop (EKAW'97), Sant Feliu de Guixols, Spain*, pages 365–370. Springer-Verlag, oct. 1997.

[Rad89] Rada R., Mili H., Bicknell E. & Bletter M. Development and application of a metric on semantic nets. In *IEEE Transaction on systems, Man and Cybernetics*, volume 19-1, Jan., Feb. 1989.

[Tes88] Tesnière L. *Élements de syntaxe structurale.* Éditions Klincksieck, Paris, 2^{me} édition, 1988.

[Tho95] Thompson C. A. 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.

[Tou97] Toussaint Y., Royaute J., Muller C. & Polanco X. Analyse linguistique et infométrique pour l'acquisition et la structuration de connaissances. In Université Toulouse-le Mirail Équipe de Recherche en Syntaxe et Sémantique, editor, *Actes des deuxièmes rencontres terminologiques et intelligence artificielle (TIA'97)*, pages 27–45, Université Toulouse-le Mirail, Avril 1997.

[Wil97] Wilks Y. Information Extraction as a Core Language Technology. In Maria Teresa Pazienza, editor, *Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology*, Frascati, Italy, July, 14-18 1997. LNAI Tutorial, Springer.

[Zel93] Zelle J. M., Mooney R. J. Learning semantic grammars with constructive inductive logic programming. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, pages 817–822. Washigton D.C., 1993.