Selection Restrictions Acquisition from Corpora

Pablo Gamallo*, Alexandre Agustini **, and Gabriel P. Lopes

CENTRIA, Departamento de Informática Universidade Nova de Lisboa, Portgual
{gamallo,aagustini,gpl}@di.fct.unl.pt

© Springer-Verlag

Abstract. This paper describes an automatic clustering strategy for acquiring selection restrictions. We use a knowledge-poor method merely based on word cooccurrence within basic syntactic constructions; hence, neither semantic tagged corpora nor man-made lexical resources are needed for generalising semantic restrictions. Our strategy relies on two basic linguistic assumptions. First, we assume that two syntactically related words impose semantic selectional restrictions to each other (*co-specification*). Second, it is also claimed that two syntactic contexts impose the same selection restrictions if they cooccur with the same words (*contextual hypothesis*). In order to test our learning method, preliminary experiments have been performed on a Portuguese corpus.

1 Introduction

The general aim of this paper is to describe a particular corpus-based method for semantic information extraction. More precisely, we implement a knowledge-poor system that uses syntactic information to acquire selection restrictions and semantic preferences constraining word combination.

According to Gregory Grefenstette [9, 10], knowledge-poor approaches use no presupposed semantic knowledge for automatically extracting semantic information. They are characterised as follows: no domain-specific information is available, no semantic tagging is used, and no static sources as machine readable dictionaries or handcrafted thesauri are required. Hence, they differ from knowledge-rich approaches in the amount of linguistic knowledge they need to activate the semantic acquisition process. Whereas knowledge-rich approaches require previously encoded semantic information (semantic tagged corpora and/or man-made lexical resources [17, 6, 1]), knowledge-poor methods only need a coarse-grained notion of linguistic information: word cooccurrence. In particular, the main aim of knowledge-poor approaches is to calculate the frequency of word cooccurrences within either syntactic constructions or sequences of n-grams in order to extract semantic information such as selection restrictions [19, 11, 3], and word ontologies [12, 15, 9, 13]. Since these methods do not require previously defined

* Research supported by the PRAXIS XXI project, FCT/MCT, Portugal
** Research sponsored by CAPES and PUCRS - Brazil

semantic knowledge, they overcome the well-known drawbacks associated with handcrafted thesauri and supervised strategies.

Nevertheless, our method differs from standard knowledge-poor strategies on two specific issues: both the way of extracting word similarity and the way of defining syntactic contexts. We use the contextual hypothesis to characterise word similarity and the co-specification hypothesis to define syntactic contexts.

1.1 The Contextual Hypothesis

In most knowledge-poor approaches to selection restriction learning, the process of inducing and generalising semantic information from word cooccurrence frequencies consists in automatically clustering words considered as similar. The best-known strategy for measuring word similarity is based on Harris' *distributional hypothesis*. According to this assumption, words cooccurring in similar syntactic contexts are semantically similar and, then, should be clustered into the same semantic class. However, the learning methods based on the distributional hypothesis may lead to cluster in the same class words that fill different selection restrictions. Let's analyse the following examples taken from [20]:

(a) John worked till late at the *council*
(b) John worked till late at the *office*
(c) the *council* stated that they would raise taxes
(d) the *mayor* stated that he would raise taxes

On the basis of the distributional hypothesis, since *council* behaves similarly to *office* and *mayor* they would be clustered together into the same word class. Nevertheless, the bases for the similarity between *council* and *office* are different from those relating *council* and *mayor*. Whereas *council* shares with *office* syntactic contexts associated mainly with LOCATIONS (e.g., the argument of *work at* in phrases (a) and (b)), *council* shares with *mayor* contexts associated with AGENTS (e.g., the subject of *state* in phrases (c) and (d)). That means that a polysemous word like *council* should be clustered into various semantic word classes, according to its heterogeneous syntactic distribution. Each particular sense of the word is related to a specific type of distribution. Given that the clustering methods based on the distributional hypothesis solely take into account the global distribution of a word, they are not able to separate and acquire its different contextual senses.

In order to extract contextual word classes from the appropriate syntactic constructions, we claim that similar syntactic contexts share the same semantic restrictions on words. Instead of computing word similarity on the basis of the too coarse-grained distributional hypothesis, we measure the similarity between syntactic contexts in order to identify common selection restrictions. More precisely, we assume that two syntactic contexts occurring with (almost) the same words are similar and, then, impose the same semantic restrictions on those words. That is what we call *contextual hypothesis*. Semantic extraction strategies based on the contextual hypothesis may account for the semantic variance

of words in different syntactic contexts. Since these strategies are concerned with the extraction of semantic similarities between syntactic contexts, words will be clustered with regard to their specific syntactic distribution. Such clusters represent context-dependent semantic classes. Except the cooperative system Asium introduced in [5, 4, 2], few or no research on semantic extraction have been based on such a hypothesis.

1.2 Co-Specification

Traditionally, a binary syntactic relationship is constituted by both the word that imposes linguistic constraints (the predicate) and the word that must fill such constraints (its argument). In a syntactic relationship, each word plays a fixed role. The argument is perceived as the word specifying or modifying the syntactic-semantic constraints imposed by predicate, while the latter is viewed as the word specified or modified by the former. However, recent linguistic research assumes that the two words related by a syntactic dependency are mutually specified [16, 7]. Each word imposes semantic conditions on the other word of the dependency, and each word elaborates them. Consider the relationship between the polysemic verb *load* and the polysemic noun books in the non ambiguous expression *to load the books*. On the one hand, the polysemic verb *load* conveys at least two alternate meanings: "bringing something to a location" (e.g., *Ann loaded the hay onto the truck*), and "modifying a location with something" (e.g., *Ann loaded the truck with the hay*). This verb is disambiguated by taking into account the sense of the words with which it combines within the sentence. On the other hand, the noun *book(s)* is also a polysemic expression. Indeed, it refers to different types of entities: "physical objects" (*rectangular book*), and "symbolic entities" (*interesting book*). Yet, the constraints imposed by the words with which it combines allows the noun to be disambiguated. Whereas the adjective *rectangular* activates the physical sense of *book*, the adjective *interesting* makes reference to its symbolic content.

In *to load the books*, the verb *load* activates the physical sense of the noun, while *books* leads *load* to refer to the event of bringing something to a location. The interpretation of the complex expression is no more ambiguous. Both expressions, *load* and *books*, cooperate to mutually restrict their meaning. The process of mutual restriction between two related words is called by Pustejovsky "co-specification" or "co-composition" [16]. Co-specification is based on the following idea. Two syntactically dependent expressions are no longer interpreted as a standard pair "predicate-argument", where the predicate is the active function imposing the semantic preferences on a passive argument, which matches such preferences. On the contrary, each word of a binary dependency is perceived simultaneously as a predicate and an argument. That is, each word both imposes semantic restrictions and matches semantic requirements. When one word is interpreted as an active functor, the other is perceived as a passive argument, and conversely. Both dependent expressions are simultaneously active and passive compositional terms. Unlike most work on selection restrictions learning, our notion of "predicate-argument" frame relies on the active process of semantic

co-specification, and not on the trivial operation of argument specification. This trivial operation only permits the one-way specification and disambiguation of the argument by taking into account the sense of the predicate. Specification and disambiguation of the predicate by the argument is not considered.

In this paper, we describe a knowledge-poor unsupervised method for acquiring selection restrictions, which is based on the contextual and the co-specification hypotheses. The different parts of this method will be outlined in the next section.

2 System Overview

To evaluate the hypotheses presented above, a software package was developed to support the automatic acquisition of semantic restrictions. The system is constituted by four related modules, illustrated in Figure 1. In the following paragraphs we merely outline the overall functionalities of these modules. Then, in the remainder of the paper, we describe accurately the specific objects and processes of each module.

Parsing: The raw text is tagged [14] and partially analysed [18]. Then, an attachment heuristic is used to identify *binary dependencies*. The result is a list of cooccurrence triplets containing the syntactic relationship and the lemmas of the two related head words. This module will be described in section 3.1

Extracting: The binary dependencies are used to extract the *syntactic contexts*. Unlike most work on selection restrictions learning, the characterisation of syntactic contexts relies on the dynamic process of co-specification. Then, the word sets that appear in those contexts are also extracted. The result is a list of *contextual word sets*. This module will be analysed in section 3.2.

Filtering: Each pair of contextual word sets are statistically compared using a variation of the weighted Jaccard Measure [9]. For each pair of contextual sets considered as similar, we select only the words that they share. The result is a list of semantically homogenous word sets, called *basic classes*. Section 4.1 describes this module.

Clustering: Basic classes are successively aggregated by a conceptual clustering method to induce more general classes, which represent extensionally the selection restrictions of syntactic contexts. We present this module in section 4.2. Finally, the classes obtained by clustering are used to update the subcategorisation information in the dictionary.[1]

The system was tested over the Portuguese text corpora *P.G.R.*[2]. Some results are analysed in section 4.3. The fact of using specialised text corpora makes

[1] The first tasks, tagging and parsing, are based on a non domain-specific dictionary for Portuguese, which merely contains morphosyntactic information. The subcategorisation restrictions extracted by our method are used to extend the morphosyntactic information of the dictionary.

[2] P.G.R. (*Portuguese General Attorney Opinions*) is constituted by case-law documents.

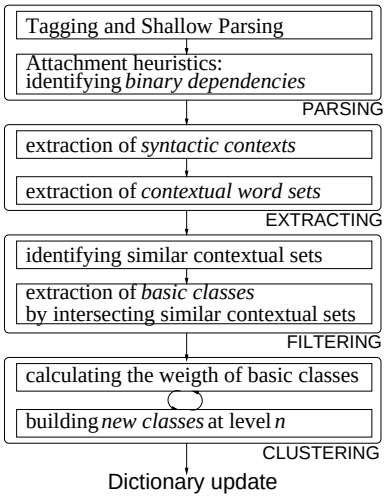

Fig. 1. System modules

easier the learning task, given that we have to deal with a limited vocabulary with reduced polysemy. Furthermore, since the system is not dependent of a specific language such as Portuguese, it could be applied to whatever natural language.

3 Identification of Binary Dependencies and Extraction of Syntactic Contexts

Binary dependencies and syntactic contexts are directly associated with the notion of selection restrictions. Selection restrictions are the semantic constraints that a word needs to match in order to be syntactically dependent and attached to another word. According to the co-specification hypothesis, two dependent words can be analysed as two syntactic contexts of specification. So, before describing how selection restrictions are learned, we start by defining first how binary dependencies are identified, and second how syntactic contexts are extracted from binary dependencies.

3.1 Binary Dependencies

We assume that basic syntactic contexts are extracted from binary syntactic dependencies. We use both a shallow syntactic parser and a particular attachment heuristic to identify binary dependencies. The parser produces a single partial syntactic description of sentences, which are analysed as sequences of basic chunks (NP, PP, VP, ...). Then, attachment is temporarily resolved by a simple heuristic based on right association (a chunk tend to attach to another chunk immediately to its right). Finally, we consider that the word heads of two

attached chunks form a binary dependency. It can be easily seen that syntactic errors may appear since the attachment heuristic does not take into account distant dependencies.[3] For reasons of attachment errors, it is argued here that the binary dependencies identified by our basic heuristic are mere hypotheses on attachment; hence they are mere candidate dependencies. Candidate dependencies will be checked by taking into account further information on the attached words. In particular, a candidate dependency will be checked and finally verified only if the two attached words impose selection restrictions to each other. Therefore, the test confirming or not a particular attachment relies on the semantic information associated with the related words. Let's describe first the internal structure of a candidate dependency between two words.

A candidate syntactic dependency consists of two words and the hypothetical grammatical relationship between them. We represent a dependency as the following binary predication: $(r; w1^{\downarrow}, w2^{\uparrow})$. This binary predication is constituted by the following entities:

- the binary predicate r, wich can be associated to specific prepositions, subject relations, direct object relations, etc.;
- the roles of the predicate, "$\downarrow$" and "$\uparrow$", which represent the *head* and *complement* roles, respectively;
- the two words holding the binary relation: $w1$ and $w2$.

Binary dependencies denote grammatical relationships between the head and its complement. The word indexed by "$\downarrow$" plays the role of *head*, whereas the word indexed by "$\uparrow$" plays the role of *complement*. Therefore, $w1$ is perceived as the head and $w2$ as the complement.

Furthermore, the binary dependencies (i.e., grammatical relationships) we have considered are the following: subject (noted *subj*), direct object (noted *dobj*), prepositional object of verbs, and prepositional object of nouns, both noted by the specific preposition.

3.2 Extraction of Syntactic Contexts and Co-Specification

Syntactic contexts are abstract configurations of specific binary dependencies. We use λ-abstraction to represent the extraction of syntactic contexts. A syntactic context is extracted by λ-abstracting one of the related words of a binary dependency. Thus, two complementary syntactic context can be λ-abstracted from the binary predication associated with a syntactic dependency: $[\lambda x^{\downarrow}(r; x^{\downarrow}, w2^{\uparrow})]$ and $[\lambda x^{\uparrow}(r; w1^{\downarrow}, x^{\uparrow})]$.

The syntactic context of word w2, $[\lambda x^{\downarrow}(r; x^{\downarrow}, w2^{\uparrow})]$, can be defined extensionally as the set of words that are the *head* of $w2$. The exhaustive enumeration of every word that can occur with that syntactic frame enables us to characterise extensionally its selection restrictions. Similarly, The syntactic context of word

[3] The errors are caused, not only by the too restrictive attachment heuristic, but also by further misleadings, e.g., words missing from the dictionary, words incorrectly tagged, other sorts of parser limitations, etc.

$w1$, $[\lambda x^{\uparrow}(r; w1^{\downarrow}, x^{\uparrow})]$, represents the set of words that are *complement* of $w1$. This set is perceived as the extensional definition of the selection restrictions imposed by the syntactic context. Consider Table 1. The left column contains expressions constituted by two words syntactically related by a particular type of syntactic dependency. The right column contains the syntactic contexts extracted from these expressions. For instance, from the expression `presidente da república` (*president of the republic*), we extract two syntactic contexts: both $[\lambda x^{\downarrow}(de; x^{\downarrow}, república^{\uparrow})]$, where `república` plays the role of *complement*, and $[\lambda x^{\uparrow}(de; presidente^{\downarrow}, x^{\uparrow})]$, where `presidente` is the *head*.

Table 1. Syntactic contexts extracted from binary expressions

Binary Expressions	Syntactic Contexts
`presidente da república` (*president of the republic*)	$[\lambda x^{\downarrow}(de;x^{\downarrow},república^{\uparrow})]$, $[\lambda x^{\uparrow}(de;presidente^{\downarrow},x^{\uparrow})]$
`nomeação do presidente` (*nomination for president*)	$[\lambda x^{\downarrow}(de;x^{\downarrow},presidente^{\uparrow})]$, $[\lambda x^{\uparrow}(de;nomeação^{\downarrow},x^{\uparrow})]$
`nomeou o presidente` (*nominated the president*)	$[\lambda x^{\downarrow}(dobj;x^{\downarrow},presidente^{\uparrow})]$, $[\lambda x^{\uparrow}(dobj;nomear^{\downarrow},x^{\uparrow})]$
`discutiu sobre a nomeação` (*disscussed about the nomination*)	$[\lambda x^{\downarrow}(sobre;x^{\downarrow},nomeação^{\uparrow})]$, $[\lambda x^{\uparrow}(sobre;discutir^{\downarrow},x^{\uparrow})]$

Since syntactic configurations impose specific selectional preferences on words, the words that match the semantic preferences (or selection restrictions) required by a syntactic context should constitute a semantically homogeneous word class. Consider the two contexts extracted from `presidente da república`. On the one hand, context $[\lambda x^{\uparrow}(de; presidente^{\downarrow}, x^{\uparrow})]$ requires a particular noun class, namely human organizations. In corpus P.G.R., this syntactic context selects for nouns such as `república` (*republic*), `governo` (*government*), `instituto` (*institute*), conselho (*council*),... On the other hand, context $[\lambda x^{\downarrow}(r; x^{\downarrow}, república^{\uparrow})]$ requires nouns denoting either human beings or organizations: `presidente` (*president*), `ministro` (*minister of state*), `assembleia` (*assembly*), `governo`, (*government*) `procurador` (*attorney*), `procuradoria-geral` (*attorneyship*) , `ministério` (*state department*), etc.

It follows that the two words related by a syntactic dependency are mutually determined. The context constituted by a word and a specific function imposes semantic conditions on the other word of the dependency. The converse is also true. As has been said, the process of mutual restriction between two related words is called co-specification. In `presidente da república`, the context constituted by the noun `presidente` and the grammatical function *head* somehow restricts the sense of `república`. Conversely, both the noun `república` and the role of *complement* also restrict the sense of `presidente`:

- $[\lambda x^{\downarrow}(de; x^{\downarrow}, república^{\uparrow})]$ selects for `presidente`
- $[\lambda x^{\uparrow}(de; presidente^{\downarrow}, x^{\uparrow})]$ selects for `república`

In our system, the extraction module consists of the two following tasks: first, the syntactic contexts associated with all candidate binary dependencies are extracted. Then, the set of words appearing in those syntactic contexts are selected. The words appearing in a particular syntactic context form a *contextual word set*. Contextual word sets are taken as the input of the processes of filtering and clustering; these processes will be described in the next section.

Let's note finally that unlike the Grefenstette's approach [9], information on co-specification is available for the characterisation of syntactic contexts. In [8], a strategy for measuring word similarity based on the co-specification hypothesis was compared to the Grefensetette's strategy. Experimental tests demonstrated that co-specification allows a finer-grained characterisation of syntactic contexts.

4 Filtering and Clustering

According to the contextual hypothesis introduced above, two syntactic contexts that select for the same words should have the same extensional definition and, then, the same selection restrictions. So, if two contextual word sets are considered as similar, we infer that their associated syntactic contexts are semantically similar and share the same selection restrictions. In addition, we also infer that these contextual word sets are semantically homogeneous and represent a contextually determined class of words. Let's take the two following syntactic contexts and their associated contextual word sets:

$$[\lambda x^{\uparrow}(of; infringement^{\downarrow}, x^{\uparrow})] = \{article\ law\ norm\ precept\ statute\ \ldots\}$$
$$[\lambda x^{\uparrow}(dobj; infringe^{\downarrow}, x^{\uparrow})] = \{article\ law\ norm\ principle\ right\ \ldots\}$$

Since both contexts share a significant number of words, it can be argued that they share the same selection restrictions. Furthermore, it can be inferred that their associated contextual sets represent the same context-dependent semantic class. In our corpus, context $[\lambda x^{\uparrow}(dobj; violar^{\downarrow}, x^{\uparrow})]$ (*to infringe*) is not only considered as similar to context $[\lambda x^{\downarrow}(dobj; violação^{\downarrow}, x^{\uparrow})]$ (*infringement of*), but also to other contexts such as: $[\lambda x^{\downarrow}(dobj; respeitar^{\downarrow}, x^{\uparrow})]$ (*to respect*) and $[\lambda x^{\uparrow}(dobj; aplicar^{\downarrow}, x^{\uparrow})]$ (*to apply*).

In this section, we will specify the procedure for learning context-dependent semantic classes from the previously extracted contextual sets. This will be done in two steps:

- Filtering: word sets are automatically cleaned by removing those words that are not semantically homogenous.
- Conceptual clustering: the previously cleaned sets are successively aggregated into more general clusters. This allows us to build more abstract semantic classes and, then, to induce more general selection restrictions.

4.1 Filtering

As has been said in the introduction, the cooperative system Asium is also based on the contextual hypothesis [5, 4]. This system requires the interactive participation of a language specialist in order to filter and clean the word sets when

they are taken as input of the clustering strategy. Such a cooperative method proposes to manually remove from the sets those words that have been incorrectly tagged or analysed. Our strategy, by contrast, intends to automatically remove incorrect words from sets. Automatic filtering consists of the following subtasks:

First, each word set is associated with a list of its most similar sets. Intuitively, two sets are considered as similar if they share a significant number of words. Various similarity measure coefficients were tested to create lists of similar sets. The best results were achieved using a particular weighted version of the Jaccard coefficient, where words are weighted considering their dispersion (global weight) and their relative frequency for each context (local weight). Word dispersion (global weight) *disp* takes into account how many different contexts are associated with a given word and the word frequency in the corpus. The local weight is calculated by the relative frequency fr of the pair word/context. The weight of a word with a context *cntx* is computed by the following formula:

$$W(word_i, cntx_j) = log_2(fr_{ij}) * log_2(disp_i)$$

where

$$fr_{ij} = \frac{frequency\, of\, word_i\, with\, cntx_j}{sum\, of\, frequencies\, of\, words\, occurring\, in\, cntx_j}$$

and

$$disp_i = \frac{\sum_j frequency\, of\, word_i\, with\, cntx_j}{number\, of\, contexts\, with\, word_i}$$

So, the weighted Jaccard similarity WJ between two contexts m and n is computed by[4]:

$$WJ(cntx_m, cntx_n) = \frac{\sum_{common_i}(W(cntx_m, word_i) + W(cntx_n, word_i))}{\sum_j (W(cntx_m, word_j) + W(cntx_n, word_j))}$$

Then, once each contextual set has been compared to the other sets, we select the words shared by each pair of similar sets, i.e., we select the intersection between each pair of sets considered as similar. Since words that are not shared by two similar sets could be incorrect words, we remove them. Intersection allows us to clear sets of words that are not semantically homogenous. Thus, the intersection of two similar sets represents a semantically homogeneous class, which we call *basic class*. Let's take an example. In our corpus, the most similar set to $[\lambda x^{\uparrow}(de; violação^{\downarrow}, x^{\uparrow})]$ (*infringement of*)) is $[\lambda x^{\uparrow}(dobj; violar^{\downarrow}, x^{\uparrow})]$ (*infringe*) . Both sets share the following words:

`sigilo princípios preceito plano norma lei estatuto disposto disposição direito convenção artigo`
(secret principle precept plan norm law statute rule precept right convention article)

[4] *common* means that just common words to both contexts m and n are computed

This basic class does not contain incorret words such as `vez, flagrantemente, obrigação, interesse` (*time, notoriously, obligation, interest*), which were oddly associated to the context $[\lambda x^{\uparrow}(de; violação^{\downarrow}, x^{\uparrow})]$, but which does not appear in context $[\lambda x^{\uparrow}(dobj; violar^{\downarrow}, x^{\uparrow})]$. This class seems to be semantically homogenous because it contains only words referring to legal documents. Once basic classes have been created, they are used by the conceptual clustering algorithm to build more general classes. Note that this strategy do not remove neither infrequent nor very frequent words. Frequent and infrequent words may be semantic significant provided that they occur with similar syntactic contexts.

4.2 Conceptual Clustering

We use an agglomerative (bottom-up) clustering for successivelly aggregating the previously created basic classes. Unlike most research on conceptual clustering, aggregation does not rely on a statistical distance between classes, but on empirically set conditions and constraints [21]. These conditions will be discussed below. Figure 2 shows two basic classes associated with two pairs of similar syntactic contexts. $[CONTX_i]$ represents a pair of syntactic contexts sharing the words `preceito, lei, norma` (*precept, law, norm*, and $[CONTX_j]$ represents a pair of syntactic contexts sharing the words `preceito, lei, direito` (*precept, law, right*). Both basic classes are obtained from the filtering process described in the previous section. Figure 3 illustrates how basic classes are aggregated into more general clusters. If two classes fill the conditions that we will define later, they can be merged into a new class. The two basic classes of the example are clustered into the more general class constituted by `preceito, lei, norma, direito`. Such a generalisation leads us to induce syntactic data that does not appear in the corpus. Indeed, we induce both that the word `norma` may appear in the syntactic contexts represented by $[CONTX_j]$, and that the word `direito` may be attached to the syntactic contexts represented by $[CONTX_i]$.

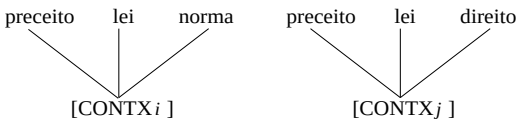

Fig. 2. Basic classes

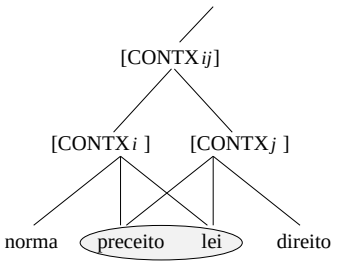

Fig. 3. Agglomerative clustering

Two basic classes are compared and then aggregated into a new more general class if they fulfil three specific conditions:

1. They must have the same number of words. We consider that two classes are compared in a more efficient manner when they have the same number of elements. Indeed, nonsensical results could be obtained if we compare large classes, which still remain polysemic and then heterogeneous, to the small classes that are included in them.
2. They must share $n-1$ words. Two classes sharing $n-1$ words are aggregated into a new class of $n+1$ members. Indeed, two classes with the same number of elements only differing in one word may be considered as semantically close.
3. They must have the highest weight. The weight of a class corresponds to the number of occurrences of the class as a subset of other classes (within $n+20$ supersets). Intuitively, the more a class is included in larger classes, the more semantically homogeneous it should be. Only those classes with the highest weight will be compared and aggregated.

Note that clustering does not rely here on a statistical distance between classes. Rather, clustering is guided by a set of constraints, which have been empirically defined considering linguistic data. Due to the nature of these constraints, the clustering process should start with small size classes with n elements, in order to create larger classes of $n+1$ members. All classes of size n that fulfil the conditions stated above are aggregated into $n+1$ clusters. In this agglomerative clustering strategy, level n is defined by the classes with n elements. The algorithm continues merging clusters at more complex levels and stops when there are no more clusters fulfilling the three conditions.

4.3 Tests and Evaluation

We used a small corpus with 1,643,579 word occurrences, selected from the case-law P.G.R. text corpora. First, the corpus was tagged by the part-of- speech tagger presented in [14]. Then, it was analysed in sequences of basic chunks by the partial parser presented in [18]. The chunks were attached using the right association heuristic so as to create binary dependencies. 211,976 different syntactic contexts with their associated word sets were extracted from these dependencies. Then, we filter these contextual word sets by using the method described above so as to obtain a list of basic classes.

In order to test our clustering strategy, we start the algorithm with basic classes of size 4 (i.e., classes with 4 elements). We have $7,571$ basic classes with 4 elements, but only a small part of them fills the clustering conditions so as to form $1,243$ clusters with 5 elements. At level 7, there are still 600 classes filling the clustering conditions, 263 at level 9, 112 at level 11, 38 at level 13, and finally only 1 at level 19. In table 2, we show some of the clusters generated by the algorithm at different intermediate levels.[5]

Note that some words may appear in different clusters. For instance, `cargo` (*task/post*) is associated with nouns referring to activities (e.g., `actividade,`

[5] In the left column, the first number represents the weight of the set, i.e., its occurrences as subset of larger supersets; the second number represents class cardinality.

Table 2. Some clusters at levels 6, 7, 8, 9, 10, and 11

0006 (06)	aludir citar enunciar indicar mencionar referir *allude cite enunciate indicate mention refer*
0009 (07)	considerar constituir criar definir determinar integrar referir *consider constitute create define determinate integrate refer*
0002 (07)	actividade atribuição cargo função funções tarefa trabalho *activity attribution position/task function functions task work*
0003 (08)	administração cargo categoria exercício função lugar regime serviço *administration post rank practice function place regime service*
0002 (09)	abono indemnização multa pensão propina remuneração renda sanção vencimento *bail compensation fine pension fee remuneration rent sanction salary*
0007 (10)	administração autoridade comissão conselho direcção estado governo ministro tribunal órgão *administration authority commission council direction state government minister tribunal organ*
0026 (11)	alínea artigo código decreto diploma disposição estatuto legislação lei norma regulamento *paragraph article code decret diploma disposition statute legislation law norm regulation*

`trabalho, tarefa` (*activity, work, task*)), as well as with nouns referring to the positions where those activities are produced (e.g., `cargo, categoria, lugar` (*post, rank, place*)). The sense of polysemic words is represented by the natural attribution of a word to various clusters.

Table 3. Some clusters generated at level 12

0002 (12)	administração associação autoridade comissão conselho direcção entidade estado governo ministro tribunal órgão *administration association authority commission council direction entity state government minister government tribunal organ*
0002 (12)	administração assembleia autoridade comissão conselho direcção director estado governo ministro tribunal órgão *administration assembly authority commission council direction director state government minister government tribunal organ*
0002 (12)	assembleia autoridade câmara comissão direcção estado europol governo ministério pessoa serviço órgão *assembly authority chamber commission direction state europol government state_department person service organ*
0002 (12)	administração autoridade comissão conselho direcção empresa estado gestão governo ministério serviço órgão *administration authority commission council direction firm state management gouvernment state_department person service organ*

Since this clustering strategy have been conceived to assure the semantic homogeneity of clusters, it does not really assure that each cluster represents an independent semantic class. Hence, two or more clusters can represent the same contextual-based class and, then, the same semantic restriction. Let's see

Table 3. Intuitively, the four clusters, which have been generated at level 12, refer to a general semantic class: *agentive_entities*. Yet, the algorithm is not able to aggregate them into a more general cluster at level 13, since they do not fill condition 2. Further work should refine the algorithm in order to solve such a problem.

Note that the algorithm does not generate ontological classes like *human beings, institutions, vegetables, dogs*,... but context-based semantic classes associated with syntactic contexts. Indeed, the generated clusters are not linguistic-independent objects but semantic restrictions taking part in the syntactic analysis of sentences. This way, the words `autoridade, pessoa, administração`, etc. (*authority, person, administration*) belong to the same contextual class because they share a great number of syntactic contexts, namely they appear as the subject of verbs such as `aprovar, revogar, considerar`, ... (*approve, repeal, consider*). Those nouns do not form an ontological class but rather a linguistic class used to constrain the syntactic word combination. More precisely, we may infer that the following syntactic contexts:

$$\begin{bmatrix}\lambda x^{\uparrow}(subj; aprovar^{\downarrow}, x^{\uparrow})\end{bmatrix}$$
$$\begin{bmatrix}\lambda x^{\uparrow}(subj; revogar^{\downarrow}, x^{\uparrow})\end{bmatrix}$$
$$\begin{bmatrix}\lambda x^{\uparrow}(subj; considerar^{\downarrow}, x^{\uparrow})\end{bmatrix}$$

share the same selection restrictions since they are used to build a context-based semantic class constituted by words like `autoridade, pessoa, administração`, ... As has been said, the acquired selection restrictions should be used to check the attachment hypotheses concerning the candidate dependencies previously extracted.

5 Conclusion and Further Work

This paper has presented a particular unsupervised strategy to automatically learn context-based semantic classes used as restrictions on syntactic combinations. The strategy is mainly based on two linguistic assumptions: co-specification hypothesis, i.e., the two related expressions in a binary dependency impose semantic restrictions to each other, and contextual hypothesis, i.e., two syntactic contexts share the same semantic restrictions if they cooccur with the same words.

The system introduced in this paper will be applied to attachment resolution in syntactic analysis. The degree of efficacy in such a task may serve as a reliable evaluation for measuring the soundness of our learning strategy. This is part of our current work.

References

1. Roberto Basili, Maria Pazienza, and Paola Velardi. Hierarchical clustering of verbs. In *Workshop on Acquisition of Lexical Knowledge from Text*, pages 56–70, Ohio State University, USA, 1993.

2. Gilles Bisson, Claire Nédellec, and Dolores Canamero. Designing clustering methods for ontology building: The mo'k workbench. In *Internal rapport*, citerseer.nj.nec.com/316335.html, 2000.
3. Ido Dagan, Lillian Lee, and Fernando Pereira. Similarity-based methods of word coocurrence probabilities. *Machine Learning*, 43, 1998.
4. David Faure. *Conception de méthode d'aprentissage symbolique et automatique pour l'acquisition de cadres de sous-catégorisation de verbes et de connaissances sémantiques à partir de textes : le système ASIUM*. PhD thesis, Université Paris XI Orsay, Paris, France, 200.
5. David Faure and Claire Nédellec. Asium: Learning subcategorization frames and restrictions of selection. In *ECML98, Workshop on Text Mining*, 1998.
6. Francesc Ribas Framis. On learning more appropriate selectional restrictions. In *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, Dublin, 1995.
7. Pablo Gamallo. *Construction conceptuelle d'expressions complexes: traitement de la combinaison nom-adjectif*. PhD thesis, Université Blaise Pascal, Clermont-Ferrand, France, 1998.
8. Pablo Gamallo, Caroline Gasperin, Alexandre Agustini, and Gabriel P. Lopes. Syntactic-based methods for measuring word similarity. In *Text, Speech, and Discourse (TSD-2001)*. Berlin:Springer Verlag (to appear), 2001.
9. Gregory Grefenstette. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, USA, 1994.
10. Gregory Grefenstette. Evaluation techniques for automatic semantic extraction: Comparing syntatic and window based approaches. In Branimir Boguraev and James Pustejovsky, editors, *Corpus processing for Lexical Acquisition*, pages 205–216. The MIT Press, 1995.
11. Ralph Grishman and John Sterling. Generalizing automatically generated selectional patterns. In *Proceedings of the 15th International on Computational Linguistics (COLING-94)*, 1994.
12. D. Hindle. Noun classification form predicate-argument structures. In *Proceedings of the 28th Meeting of the ACL*, pages 268–275, 1990.
13. Dekang Lin. Automatic retrieval and clustering of similar words. In *COLING-ACL'98*, Montreal, 1998.
14. Nuno Marques. *Uma Metodologia para a Modelação Estatística da Subcategorização Verbal*. PhD thesis, Universidade Nova de Lisboa, Lisboa, Portugal, 2000.
15. Fernando Pereira, Naftali Tishby, and Lillian Lee. Distributional clustering of english words. In *Proceedings of the 30th Annual Meeting of the Association of Comptutational Linguistics*, pages 183–190, Columbos, Ohio, 1993.
16. James Pustejovsky. *The Generative Lexicon*. MIT Press, Cambridge, 1995.
17. Philip Resnik. Semantic similarity in taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *Journal of Artificial Intelligence Research*, 11:95–130, 1999.
18. V. Rocio, E. de la Clergerie, and J.G.P. Lopes. Tabulation for multi-purpose partial parsing. *Journal of Grammars*, 4(1), 2001.
19. Satoshi Sekine, Jeremy Carrol, Sofia Ananiadou, and Jun'ichi Tsujii. Automatic learning for semantic collocation. In *Proceedings of the 3rd Conference on Applied Natural Language Processing*, pages 104–110, 1992.
20. Tokunaga Takenobu, Iwayama Makoto, and Tanaka Hozumi. Automatic thesaurus construction based on grammatical relations. In *Proceedings of IJCAI-95*, 1995.
21. Luis Talavera and Javier Béjar. Integrating declarative knowledge in hierarchical clustering tasks. In *Intelligent Data Analysis*, pages 211–222, 1999.