The Role of Co-Specification for the Acquisition of Selection Restrictions from Unsupervised Corpora

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

CENTRIA, Departamento de Informática Universidade Nova de Lisboa
Quinta da Torre, 2825-114 Caparica, Portugal
{gamallo,aagustini,gpl}@di.fct.unl.pt

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 training 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 frames impose the same selection restrictions if they cooccur with the same words. In order to test the learning method, preliminary experiments have been performed on a domain-specific 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 method that uses syntactic information to acquire selection restrictions or semantic preferences constraining word combination.

According to Gregory Grefenstette [7, 8], 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 to build supervised corpora, 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 training corpora and/or man-made lexical resources [15, 5, 1]), Knowledge-poor methods only need for 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 [17, 9, 2], and word ontologies [10, 13, 7, 11]. Since these methods does not require previously defined semantic knowledge, they avoid the well-known shortcomings associated with handcrafted thesauri and supervised corpora.

In most knowledge-poor methods, the process of generalising and learning semantic information from word cooccurrence frequencies relies on strategies for clustering similar words. The best-known strategy for measure word similarity is based on Harris' *distributional hypothesis*. According to this assumption, words cooccurring in similar syntactic contexts are semantic similar and, then, should be clustered into the same semantic class. However, the similarity-based 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 [18]:

(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 of the similarity between "council" and "office" are different than between "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

* This work was supported by the PRAXIS XXI project, Fundação para a Ciência e a Tecnologia, Ministério da Ciência e a Tecnologia, Portugal
** Bolsista da CAPES - Brasília/Brasil.

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 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 [4, 3], few or no research on semantic extraction have been based on such a hypothesis.

The main concern of this paper is with the contextual hypothesis. That is, we implement a computational strategy for measuring the statistical similarity among syntactic contexts. This strategy results in a knowledge-poor unsupervised method for acquiring the selection restrictions and the distributional word classes associated with syntactic contexts. We have applied this learning method on the Portuguese text corpora *P.G.R.*[1]. The fact of using specialised corpora makes easier the learning task, given that we have to deal with a limited vocabulary with reduced polysemy. The remainder of the article is organised as follows. Section 2 analyses theoretical notions such as syntactic dependency, syntactic contexts and selection restrictions. Unlike most work on selection restrictions learning, the characterisation of these notions relies on the dynamic process of co-specification [14]. Then, in section 3, we propose an automatic method for clustering the words appearing in similar syntactic frames. This method allows us to generalise contextual semantic classes on the basis of similar word sets associated with specific syntactic contexts.

2 Assumptions on Syntactic Contexts and Selectional Restrictions

Selectional restrictions are the semantic constraints that a word needs to match in order to be attached to a specific syntactic context. Syntactic contexts are linguistic configurations or specific frames that impose semantic constraints on words. So, before describing how selection restrictions are learned, we start by defining the manner in which syntactic frames are extracted from syntactic dependencies between words.

2.1 Binary Dependencies

We assume that basic syntactic contexts are extracted from binary syntactic dependencies. We use both a shallow syntactic parser [16] 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,..., where complements and adjunts are not included). 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.[2] 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. The 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:

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

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

– 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 *modified* and *modifier* roles, respectively;
– the two words holding the binary relation: w1 and w2.

Binary dependencies denote relations of grammatical modification. The word indexed by "$\downarrow$" plays the role of *modified*, whereas the word indexed by "$\uparrow$" plays the role of *modifier*. Therefore, $w1$ is modified by $w2$ as well as $w2$ modifies w1.

2.2 Syntactic Contexts and Co-specification

Syntactic contexts are abstract configurations of specific binary dependencies. A syntactic context is built from abstracting one of the related words of a binary dependency. Two complementary syntactic configurations may be λ-abstracted from the predication associated with a syntactic dependency:

$$[\lambda x^{\downarrow}(r; x^{\downarrow}, w2^{\uparrow})] \qquad [\lambda x^{\uparrow}(r; w1^{\downarrow}, x^{\uparrow})]$$

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 modified by 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 w1, $[\lambda x^{\uparrow}(r; w1^{\downarrow}, x^{\uparrow})]$, represents the set of words that modify 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 $[\lambda x^{\downarrow}(r; x^{\downarrow}, andrepública^{\uparrow})]$, where `república` plays the role of modifier word, $[\lambda x^{\uparrow}(r; presidente^{\downarrow}, x^{\uparrow})]$, where `presidente` is the modified word. Unlike the Grefenstette's approach [7], information on the specific preposition connecting the words must be available for the characterisation of syntactic contexts.

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*)	$[\lambda x^{\downarrow}(sobre; x^{\downarrow}, nomeação^{\uparrow})]$, $[\lambda x^{\uparrow}(sobre; discutir^{\downarrow}, x^{\uparrow})]$

Table 1. Syntactic contexts extracted from binary expressions

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 homogenous 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 PGR, this syntactic frame 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 humain beings or organizations: `presidente` (*president*), `ministro` (*minister of state*), `assembleia` (*assembly*), `governo`, (*government*) `procurador` (*attorney*), `procuradoria-geral` (*attorneyship*) , `ministério` (*state department*),....

This implies 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. In `presidente da república`, the frame constituted by the noun `presidente` and the syntactic function *modified* somehow restricts the sense of `república`. Conversely, both the noun `república` and the role of *modifier* also restrict the sense of `presidente`. The process of mutual restriction between two related words is called by Pustejovsky "co-specification" or "co-composition" [14, 6]. 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 that imposes the semantic preferences on a passive argument, while the argument 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 imposes semantic restrictions on the other word as well as it matches its 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 of a predetermined argument by a predetermined predicate.

2.3 The Contextual Hypothesis

According to the contextual hypothesis introduced above, two syntactic contexts that select for the same words have the same extensional definition and, then, the same selection restrictions. Those contexts that share a great number of words over the corpus should be considered as extensionally equivalent and semantically similar, hence they impose the same selection restrictions on words. In addition, those words occurring with both contexts form a semantically homogenous set representing a contextually determined word class.

In the next section, we will specify a method for measuring similar contexts and identifying semantically homogenous word sets. Then, we will use a specific clustering strategy to merge homogenous word sets into more general clusters. Such clusters represent the selection restrictions shared by similar syntactic contexts. Moreover, the learned selection restrictions also represent contextual classes that can be organised in a word ontology.

3 Pruning and Clustering

To validate the contextual hypothesis, 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 [12]. Then, it was analysed in sequences of basic chunks by the partial parser presented in [16]. The chunks were attached using the right association heuristic so as to create binary dependencies. 186,952 different syntactic contexts with their associated set words was extracted from these dependencies.

The word sets occurring with the previously extracted syntactic contexts are the input of the conceptual clustering strategy. The goal of this strategy is to learn the context-dependent semantic classes emerging from the corpus. These contextual classes represent the selection restrictions imposed by the syntactic contexts. In order to build contextual classes, we developped two statistical processes:

- Word sets pruning: 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 at different generality levels. The clusters obtained by means of this incremental method represent the contextual word classes imposed by syntactic contexts.

3.1 Pruning

As has been said above, the cooperative system Asium is also based on the contextual hypothesis [4, 3]. This systems 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 propose to manually prune and 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 pruning is developed in two steps:

First, each word set is associated with a list of its most similar sets. Intuitively, two sets are considered similar if they share a great number of words. Various similarity measure coefficients were tested to create lists of similar sets. The best results were performed by a particular weighted version of the Jaccard coefficient.

Second, we select the intersections of words between similar sets. Since words that are not shared by two similar sets are likely to be incorrect words, we remove them. The result of each intersection is a cleaned set representing a specific semantic class; we call it *basic class*. Let's take an example. In our corpus, the most similar set to $[\lambda x^{\uparrow}(de; \textit{violação}^{\downarrow}, x^{\uparrow})]$ (*infringement of*)) is $[\lambda x^{\uparrow}(odir; violar^{\downarrow}, x^{\uparrow})]$ (*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 disposition disposition right convention article)

The creation of this new word set by intersection leads to remove incorret words such as `vez, flagrantemente, estatuído, obrigação, interesse` (*time, notoriously, established, obligation, interest*), which appear into one of the sets. The result of intersecting both sets is a specific basic class. In this case the basic class represents administrative documents and abstract ideas likely to be violated. Once basic classes have been created, they are taken by the conceptual clustering algorithm in order to build more general classes.

3.2 Conceptual Clustering

We use an agglomerative (bottom-up) clustering for generalising the previously created basic classes. Two basic classes are compared and then aggregated into a new more general class if they fill 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.
2. They must share $n-1$ words; hence they are aggregated into a new class of $n+1$ members.
3. From the basic classes sharing the same $n-1$ words, only those having the highest weitht are aggregated. The weight of a class correspond to the number of occurrences of the class as a subset of other classes (within $n+20$ supersets). Intuitively, the more a class appears as a subset of other classes, the more semantically homogeneous it should be.

Due to the nature of these conditions, the clustering process should start with small size classes with n elements in order to create bigger classes of $n+1$ members. All classes of size n that fill the conditions stated above are aggregated into $n+1$ clusters. The algorithm continues merging clusters at more complexe levels and stops when there are no more clusters filling the three conditions.

3.3 Tests and Evaluation

In order to test our clustering strategy, we starts the algorithm with basic classes of size 4 (i.e., classes with 4 elements). We have $4,343$ basic classes with 4 elements, but only a small part of them fills the clustering conditions so as to form 984 clusters with 5 elements. At level 7, there are still 494 classes filling the clustering conditions, 183 at level 9, 59 at level 11, 18 at level 1, and finally only 2 at level 16. In table 2 below, we show some of the clusters generated by the algorithm at different intermidiate levels[3]:

0005 (06)	abranger considerar constituir criar definir integrar *comprise consider constitute create define integrate*
0006 (07)	carreira cargo categoria funções lugar quadro serviço *career post rank functions position staff service*
0002 (08)	crédito despesa multa receita remuneração retenção valor vencimento *credit expense fine income fee retention value salary*
0018 (09)	artigo decreto diploma direito disposição disposto lei norma regime *article decret diploma right disposition disposition law norm regime*

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

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 semantic restrictions. Let's see an example. Table 3 shows the two last clusters generated by the algorithm at level 16:

Intuitively, both clusters refers to a general semantic class: *documents*. Even though the algorithm generates these two semantically homogenous clusters, it does not allow to aggregate them into a more general cluster since they do not fill the condition 2. Let's take another example. The five clusters below, in table 4, have been generated at level 9:

These clusters seem to refer to the same general class: *agentive entities*. Yet, since they do not fill the condition 2, they cannot be aggragated into a single cluster at level 10. Further work should refine the algorithm in order to solve such a problem.

Note that the algorithm does not generate ontological concepts like *human beings, vegetables, dogs,*... but context-based semantic classes associated with syntactic contexts. Indeed, the generated clusters are not

[3] 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 the class cardinality.

0002 (16)	alínea aplicação artigo decreto diploma direito disposição disposto estatuto lei norma n ordenamento preceito regime texto *paragraph diligence article decret diploma right disposition disposition statute law norm ordinance precept regime text*
0002 (16)	alínea artigo decreto diploma direito disposição disposto estatuto lei norma n ordenamento preceito princípio princípios regime *paragraph article decret diploma right disposition disposition statute law norm ordinance precept principle principles regime*

Table 3. The two clusters generated at level 16

0002 (09)	administração autoridade entidade governo lei membro pessoa tribunal órgão *administration autority entity government law member person tribunal*
0002 (09)	administração autoridade comissão direito estado governo pessoa presidente proveedor *administration autority commission right state government person president purveyor*
0002 (09)	administração autoridade comisão conselho estado membro pessoa tribunal órgão *administration autority commission council state member person tribunal organ*
0002 (09)	administração assembleia c comisão conselho entidade estado governo instituto *administration assembly commission council entity state government institute*

Table 4. Some clusters generated at level 9

linguistic-independent objects but semantic restrictions taking part in the syntactic analysis of sentences. This way, the words `autoridade, pesoa, lei` can 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, constituir`,.... Let's note that the same word may belong to different clusters (e.g., `lei` appears as a polysemic word: it is both a *document* and an *agentive entitiy*). As has been said, selection restrictions should be used to check the attachment hypotheses concerning the candidate dependencies previously extracted.

4 Conclusion

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 first application of this work will be 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. Detecting dependencies between semantic verb subclasses and subcategorization frames in text corpora. In *Workshop on Acquisition of Lexical Knowledge from Text)*, pages 70–81, Ohio State University, USA, 1993.
2. Ido Dagan, Lillian Lee, and Fernando Pereira. Similarity-based methods of word coocurrence probabilities. *Machine Learning*, 43, 1998.
3. 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.
4. David Faure and Claire Nédellec. Asium: Learning subcategorization frames and restrictions of selection. In *10th European Conference on Machine Learning, ECML98, Workshop on Text Mining*, 1998.
5. 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.
6. Pablo Gamallo. *Construction conceptuelle d'expressions complexes: traitement de la combinaison nom-adjectif.* PhD thesis, Université Blaise Pascal, Clermont-Ferrand, France, 1998.
7. Gregory Grefenstette. *Explorations in Automatic Thesaurus Discovery.* Kluwer Academic Publishers, USA, 1994.
8. 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 Aquisition*, pages 205–216. The MIT Press, 1995.

9. Ralph Grishman and John Sterling. Generalizing automatically generated selectional patterns. In *Proceedings of the 15th International on Computational Linguistics (COLING-94)*, 1994.
10. D. Hindle. Noun classification form predicate-argument structures. In *Proceedings of the 28th Meeting of the ACL*, pages 268–275, 1990.
11. Dekang Lin. Automatic retrieval and clustering of similar words. In *COLING-ACL'98*, Montreal, 1998.
12. Nuno Marques. *Uma Metodologia para a Modelação Estatística da Subcategorização Verbal.* PhD thesis, Universidade Nova de Lisboa, Lisboa, Portugal, 2000.
13. Fernando Pereira, Naftali Tishby, and Lillian Lee. Distributional clustering of english words. In *Proceedings of the 30th Annual Meeting of the Assocition of Comptuational Linguistics*, pages 183–190, Columbos, Ohio, 1993.
14. James Pustejovsky. *The Generative Lexicon.* MIT Press, Cambridge, 1995.
15. 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.
16. V. Rocio, E. de la Clergerie, and J.G.P. Lopes. Tabulation for multi-purpose partial parsing. *Journal of Grammars*, 4(1), 2001.
17. Satoshi Sekine, Jeremy Carrol, Sofia Ananiadou, and Jun'ichi Tsujii. Automatic learning for semantic collocation. In *Proceedings of the 3th Conference on Applied Natural Language Processing*, pages 104–110, 1992.
18. Tokunaga Takenobu, Iwayama Makoto, and Tanaka Hozumi. Automatic thesaurus construction based on grammatical relations. In *Proceedings of IJCAI-95*, 1995.