Acquiring Semantic Classes to Elaborate Attachment Heuristics

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

CITI, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Portgual
{gamallo,aagustini,gpl}@di.fct.unl.pt

Abstract. We use an unsupervised method for learning word classes from partially parsed text corpora. A word class consists of those words that can appear in similar contexts of subcategorization. The main objective of this paper will be to describe an evaluation procedure based on attachment resolution. We will evaluate whether the classes that have been previously learnt are useful in that syntactic task.

1 Introduction

We use an unsupervised method for learning word classes from partially parsed text corpora. The words that occur with similar contexts of subcategorization are aggregated in semantic classes. The main objective of this paper will be to describe an evaluation procedure based on attachment resolution. Attachment resolution consists in providing a parser with information to disambiguate arcs. We will evaluate whether the classes that have been learnt previously are useful in that syntactic task. The degree of efficiency in such a task may serve as a reliable evaluation for measuring the soundness of our learning strategy.

The learning method has been accurately described in [4]. For the purpose of this paper, we will only outline the basic assumptions the acquisition strategy is based on, as well as different steps and modules that are involved in it (section 2). In section 3, two tasks are focused: first, we will describe how the learnt information is used to characterize attachment heuristics, and finally, we will evaluate the performance of these heuristics.

2 Learning Word Classes from Subcategorization Contexts

2.1 Basic Assumptions

Two theoretical assumptions underlie our method: one is based on word co-specification and the other on context similarity.

* Research supported by Program POSI, FCT/MCT, Portugal; ref: SFRH/BPD/11189/2002

** Research sponsored by CAPES and PUCRS - Brazil

First, we assume a very general notion of linguistic subcategorization. More precisely, we consider that in a Head-Modifier syntactic dependency, not only the Head imposes constraints on the Modifier, but the Modifier also imposes linguistic requirements on the Head in return. This idea stems from the Pustejovsky's "co-specification" hypothesis [10]. So, for a particular word, we attempt to learn what type of both modifiers and heads it subcategorizes. For instance, consider the compositional behavior of the noun `republic` in a domain-specific corpus. On the one hand, this word appears in the Head position within dependencies such as `republic of Ireland`, `republic of Portugal`, and so on. On the other hand, it plays the role of Modifier in dependencies like `president of the republic`, `government of the republic`, etc. So, given `republic`, we attempt to learn both what kind of complements and what kind of heads it subcategorizes. As there are interesting semantic regularities among the words co-occurring with `republic` in the above expressions, the aim of our learning method is to extract two different subcategorization contexts:

- $< of; republic, [M] >$, where position M must be filled by words referring to particular nations or states. Indeed, in semantic-pragmatic terms, only nations or states can be republics;
- $< of; [H], republic >$, where position H must be filled by head words denoting specific parts of the republic: e.g., institutions, organizations, functions, and so on.

The second assumption concerns the procedure for identifying and clustering *similar* subcategorization contexts. We assume, in particular, that different contexts are considered to be semantically similar if they have similar word distribution [3]. Let's take, for instance, the following set of contexts:

$$\begin{gathered}\{< of; [H], republic > \\ < of; [H], state > \\ < of; delegate, [M] > \\ < iobj\_by; sign, [M] > \\ < iobj\_on; be\_incumbent, [M] > \}\end{gathered} \tag{1}$$

All contexts in (1) share the same semantic preferences provided they require words denoting the same semantic class. Contexts with the same semantic preferences are likely to possess similar word distribution. Moreover, we also assume that the set of words required by similar subcategorization contexts represents the extensional description of their semantic preferences.

2.2 Method Overview

Our learning method consists in the following steps. Raw Portuguese text is automatically tagged and partially analyzed in sequences of basic chunks. Then, binary syntactic dependencies are identified on the basis of some symbolic attachment heuristics. Then, we extract subcategorization contexts from the binary dependencies, by following the first assumption outlined above (co-specification).

Finally, subcategorization contexts with similar word distributions are clustered into more general classes (second assumption). Similarity between contexts is calculated by using a particular version of the Lin coefficient [7]. For instance, the class illustrated above in (1) is constituted by contexts considered as similar. These contexts have as features those words co-occurring at least once with them, e.g.:

`president, assembly, minister, ministry, government, administration`

This is the way we build classes of words representing the semantic preferences of similar contexts. Such word classes represent contextual senses of words.

Note that this acquisition method permits to associate a polysemic word to different clusters, where each cluster represents one of its senses. For instance, word `administration` not only may appear in the cluster sketched above in (1), but also in:

$$\begin{aligned} &\{< of; [H], drugs > \\ &< of; process, [M] > \\ &< in; [H], course >\} \end{aligned} \tag{2}$$

This word does not mean here a kind of institution any more, but a type of action: `administration of drugs, process of administration`, etc.

2.3 Experiments and Results

We tested our learning strategy over a Portuguese corpus with $6,643,579$ word occurrences, selected from the *P.G.R.* (*Portuguese General Attorney Opinions*) text corpora. The data concerning the information extracted from this corpus is presented in Table 1.

Word Occurrences	6,643,579
Binary Dependencies	1,543,659
Syntactic Contexts	178,522
Final Clusters	16,274

Table 1. Corpus data

In order to evaluate the linguistic relevance of these clusters, we will check in the next section if they are useful in a parsing task. The degree of efficiency in such a task (parsing) may serve as a reliable evaluation for measuring the soundness of the learning strategy.

3 One Application: Attachment Resolution

The acquired classes are used to provide the lexicon with subcategorization information. Then, this information is used to improve the parsing task.

3.1 Lexicon Update

secretário (*secretary*)

· $< de, [H], \textit{secretário} >$ (*[H] of secretary*) =
`cargo, carreira, categoria, competência, escalão, estatuto, funȝão, remuneracaão, trabalho, vencimento`
(post, career, category, qualification, rank, status, function, remuneration, job, salary)

· $< de, \textit{secretário}, [M] >$ (*secretary of [M]*) =
`administração, assembleia, autoridade, conselho, direcção, empresa, entidade, estado, governo, instituto, juiz, ministro, ministério, presidente, serviço, tribunal órgão`
(administration, assembly, authority, council direction, company, entity, state, government , institute, judge, minister, ministery, president, service, tribunal organ)

· $< iobj\_a, [H], \textit{secretário} >$ (*[H] to the secretary*) =
`aludir, aplicar:refl, atender, atribuir, concernir, corresponder, determinar, presidir, recorrer, referir:refl, respeitar`
(allude, apply, attend, assign, concern, correspond, determine, resort, refer, relate)

· $< iobj\_a, [H], \textit{secretário} >$ (*[H] to the secretary*) =
`caber, competir, conceder, conferir, confiar, dirigir, incumbir, pertencer`
(concern, be-incombent, concede, confer, trust, send, be-incombent, belong)

· $< iobj\_por, [H], \textit{secretário} >$ (*[H] by the secretary*) =
`assinar, conceder, conferir, homologar, louvar, subscrito`
(sign, concede, confer, homologate, compliment, subscribe)

· $< subj, [H], \textit{secretário} >$ (*the secretary [H])*) =
`definir, estabelecer, fazer, fixar, indicar, prever, referir`
define, establish, make, fix, indicate, forsee, refer

Table 2. Dictionary entry of `secretário`

Table 2 shows how the acquired classes are used to provide lexical entries with information on the word classes which we have learnt previously. These classes represent subcategorization information that will be useful in the parsing procedure. The entry `secretário` (*secretary*), for instance, is associated with six subcategorization contexts: two nominal and four verbal ones. Concerning the nominal contexts, we learn that *secretary* selects for nouns such as *post*, *rank*, etc. in position H of preposition *de* (*of*), whereas it requires a class of nouns denoting institutions and functions in position M of this same preposition. Concerning the verbal contexts, we also learn that *secretary* subcategorizes various verb classes in different verbal positions: two classes in the H position of the *iobj_a* prepositional dependency, one class as H of *iobj_por*, and one more class as H of the subject relation: *subj*.

Notice that it is the notion of co-specification that allows us to acquire a great number of subcategorization contexts , which are not usual in the standard approaches to subcategorization. Five contexts of *secretary* do not subcategorize standard dependent complements, but different types of heads. This is a significant novelty of our approach.

3.2 Parsing with Co-Specification

The syntactic and semantic subcategorization information provided by the lexical entries is used to improve the parsing task. Details of a symbolic DCG grammar with information on linguistic co-specification can be found in [1, 4]. In this paper, we briefly outline how the parser uses this information to solve syntactic attachments. Co-specification is at the center of attachment resolution. It is used to characterize the main attachment heuristic. This heuristic states that two chunks are syntactically attached only if one of these two conditions is verified: either the *Modifier* is semantically required by the *Head*, or the *Head* is semantically required by the *Modifier*. Take the expression:

{1} `...compete ao secretário ...`
(*is incumbent on the secretary*)

This expression will be analyzed as a VP-PP construction if and only if, at least, one of the two following requirements is satisfied:

requirement M: context $< iobj\_a\ competir, [M] >$ (*be-incumbent on [M]*) subcategorizes a class of nouns to which `secretário` (*secretary*) belongs;
requirement H: context $< iobj\_a; [H], secretário >$ (*[H] on the secretary*) subcategorizes a class of verbs to which `competir` (*be-incumbent*) belongs.

According to the lexical information illustrated in Table 2, the expression {1} can be analyzed as a VP-PP construction because, at least, requirement H is satisfied. Note that, even if we have no information on the verb subcategorization, the attachment is allowed because of the noun requirements in the H position. Co-specification is also used to solve long-distance attachments [4].

3.3 Evaluating Performance of Attachment Resolution

We evaluated the performance of the parsing strategy based on co-specification. The general aim of this evaluation is to check whether the subcategorization information we have learnt is adequate to be used in a parsing task. The degree of efficiency in such a task may serve as a reliable evaluation for measuring the soundness of our learning strategy.

Test Data Most work on attachment resolution [5, 11, 2, 6, 8, 9] uses as test data expressions with three basic phrases (or chunks): $vp-np-pp$. These approaches consider that each expression selected for evaluation can be syntactically ambiguous in two ways. For instance, the partial parse:

{2} [$_{vp}$ cut] [$_{np}$ the potato] [$_{pp}$ with a knife]

can be disambiguated either by the parse:

{3} [$_{vp}$ cut [$_{np}$ the potato [$_{pp}$ with a knife]]]

which represents a syntactic configuration based on proximity (*phrase1* is attached to *phrase2* and *phrase2* is attached to *phrase3*), or by:

{4} [$_{vp}$ cut [$_{np}$ the potato] [$_{pp}$ with a knife]]

which is the correct configuration. It contains both a contiguous and a long distance attachment: *phrase1* is attached to *phrase2* and *phrase1* is attached to *phrase3*.

We consider, however, that the process of attachment resolution should be generalized to other syntactic sequences and ambiguity configurations. Our test data consists of 633 manually disambiguated expressions, which have been selected randomly from a test corpus. These expressions are not only $vp-np-pp$ sequences of phrases. They were divided into three groups according to three different syntactic sequences (see table 3). Moreover, they cannot be reduced to only two syntactic configurations (two parses). They can be syntactically ambiguous in several ways, including adjective and adverb associations. In table 3, the two first expressions contain adjective attachments, while the third one introduces the first part of an adverbial sentence: `na medida [em que]` (*in the sense [that]*).

$np-pp-pp$	`[`$_{np}$ `o artigo relativo] [`$_{pp}$ `ao decreto] [`$_{pp}$ `da lei]`
	(*the article referring to the decree of the law*)
$vp-pp-pp$	`[`$_{vp}$ `publicou] [`$_{pp}$ `nos estatutos anexos] [`$_{pp}$ `ao citado diploma]`
	(*published in the statutes joined to the cited diploma*)
$vp-np-pp$	`[`$_{vp}$ `tem] [`$_{np}$ `acesso] [`$_{pp}$ `na medida]`
	(*has acess in the sense*)

Table 3. The three syntactic sequences evaluated

Baseline Concerning the ability to propose correct syntactic attachments, we made a comparison between our method and a baseline strategy. As a baseline, we used the attachments proposed by Right Association. That is, for each expression of the test data, this strategy always proposes the attachment by proximity, that is: *phrase1* is attached to *phrase2*, *phrase2* is attached to *phrase3*, and *phrase1* is not attached to *phrase3*.

Precision and Recall Each expression selected from the test corpus contains three phrases and three candidate attachments. So, given a test expression, three different attachment decisions will be evaluated. The evaluation of each attachment decision taken by the system can be:

- true positive (*tp*): the system proposes a correct attachment
- true negative (*tn*): the system proposes correctly that there is no attachment.
- false positive (*fp*): the system proposes an incorrect attachment
- false negative (*fn*): the system proposes incorrectly that there is no attachment.

The evaluation test measures the ability of the system to make true decisions. We call both *tp* and *tn* "true decisions" (*td*). As far our strategy is concerned, a false negative (*fn*) is interpreted as the situation in which the system has not enough subcategorization information to make a decision. Concerning the baseline, the *fn* decisions correspond to those situations where there is a true long distance attachment: *phrase1* is attached to *phrase3*.

Taking into account these variables, *precision* is defined as the number of true decisions suggested by the system divided by the number of total suggestions. That is:

$$precision = \frac{td}{td + fp} \tag{3}$$

Recall is computed as the number of true decisions suggested by the system divided by the decisions that are actually correct:

$$recall = \frac{td}{td + fn} \tag{4}$$

Note that we do not define precision and recall in a standard way. Unlike most related work, precision and recall are defined here with respect to the decisions made by the system given a candidate attachment, and not given the whole configuration of the expression. This more accurate evaluation gives us more significant information on the system capabilities concerning attachment resolution.

Results Table 4 reports the test scores concerning the precision and recall of the two comparative experiments performed. The total precision of our method reaches more than 90%, whereas the total recall is about 75%. These results can be hardly compared to related approaches given that: i) there is no related work on Portuguese; ii) our test corpus is not restricted to the two standard ambiguity configurations defined above; we also take into account expressions containing adjective attachments and adverbial phrases; iii) we use three types of phrase sequences, and not only the $vp - np - pp$ sequence used by most related work; iv) the precision and recall definitions are not the standard ones.

This makes it difficult to compare the performance of our method to other unsupervised strategies. We consider, however, that the recall we have obtained should be higher. In order to improve it, we need to provide the dictionary with more items of subcategorization information. One of the main challenges of our current work is to tune the clustering constraints so as to reach high recall by keeping a reasonable precision. This should make the parser more efficient concerning the attachment resolution task.

BASELINE			
Syntactic sequences	**Precision** (%)	**Recall** (%)	**F-Score** (%)
$np - pp - pp$	70.81	78.93	74.65
$vp - pp - pp$	71.90	77.83	74.75
$vp - np - pp$	75.49	79.22	77.31
Total	72.74	78.66	75.58
CO-SPECIFICATION			
Syntactic sequences	**Precision** (%)	**Recall** (%)	**F-Score** (%)
$np - pp - pp$	87.15	74.13	80.11
$vp - pp - pp$	92.23	70.36	79.82
$vp - np - pp$	94.23	76.36	84.36
Total	91.20	73.62	81.47

Table 4. Evaluation of Attachment Resolution

4 Conclusion and Future Work

This paper has presented a particular unsupervised strategy to automatically acquire syntactic and semantic subcategorization requirements. Our strategy is mainly based on two linguistic assumptions: First, it was assumed that not only the syntactic Head imposes restrictions on its Dependent word, but also the latter selects for a specific type of Head. This phenomenon was called "co-specification". Second, we claimed that similar syntactic contexts share the same selection requirements. So, we measured, not similarity between words on the basis of their syntactic distribution, but similarity between syntactic contexts on the basis of their word distribution. It was assumed that the latter kind of similarity conveys more pertinent information on linguistic subcategorization than the former one. The learning process allowed us to provide the lexicon with both syntactic and semantic subcategorization information. This information was used to constrain attachment heuristics.

In future work, our aim is to extend the subcategorized lexicon in order to increase the coverage of the parser. For this purpose, we will use the partial results of the parser to discover new subcategorization information. That is, the new long distance attachments identified by the parser will serve to learn more syntactic and semantic restrictions. The lexicon will be provided with these new restrictions and thereby the coverage of the parser will be increased. The successive "learning + parsing" cycles will stop as no more new information is acquired and no more new dependencies are proposed.

References

1. Alexandre Agustini, Pablo Gamallo, and Gabriel P. Lopes. Selection restrictions acquisition for parsing improvement. In Oskar Bartenstein et alii, editor, *Web Knowledge Management and Decision Support*. Berlin:Springer Verlag, 2003.
2. Michael Collins and James Brooks. Prepositional phrase attachment through a backed-off model. In *Proceedings of the Third Workshop on Very Large Corpora*, pages 27–38, Cambridge, 1995.
3. David Faure and Claire Nédellec. Asium: Learning subcategorization frames and restrictions of selection. In *ECML98, Workshop on Text Mining*, 1998.
4. Pablo Gamallo, A. Agustini, and Gabriel P. Lopes. Learning subcategorisation information to model a grammar with co-restrictions. *Traitement Automatic de la Langue*, 44(1):93–117, 2003.
5. Donald Hindle and Mats Rooth. Structural ambiguity and lexical relations. *Computational Linguistics*, 19(1):103–120, 1993.
6. Hang Li and Naoki Abe. Word clustering and disambiguation based on co-occurrence data. In *Coling-ACL'98*, pages 749–755, 1998.
7. Dekang Lin. Automatic retrieval and clustering of similar words. In *COLING-ACL'98*, Montreal, 1998.
8. Michael Niemann. Determining pp attachment through semantic associations and preferences. In *ANLP Post Graduate Workshop*, pages 25–32, 1998.
9. Patrick Pantel and Dekan Lin. An unsupervised approach to prepositional phrase attachment using contextually similar words. In *ACL'00*, pages 101–108, Hong Kong, 2000.
10. James Pustejovsky. *The Generative Lexicon*. MIT Press, Cambridge, 1995.
11. Adwait Ratnaparkhi, Jeff Reymar, and Salim Roukos. A maximum entropy model for prepositional phrase attachment. In *Proceedings of the ARPA Human Language Technology Workshop)*, pages 250–225, 1994.