Extraction of Lexico-Semantic Classes from Text - Extended Abstract -

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

[1] Faculdade de Filologia, Universidade de Santiago de Compostela, Spain
pablogam@usc.es
[2] Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Portgual
gpl@di.fct.unl.pt
[3] Departamento de Informática, PUCRS, Brazil
agustini@inf.pucrs.br

1 Introduction

This paper describes an unsupervised method for extracting lexico-semantic classes from *POS* annotated corpora. The method consists in building bi-dimensional clusters of both words and local syntactic contexts. Each cluster, which represents a lexico-semantic class such as "entities in danger" is the result of merging its most prototypical constituents (words and contexts). The generated clusters will be used as centroids to word classification.

The basic intuition underlying our corpus-based approach is that similar classes can be aggregated to generate either more specific or more generic classes, without inducing odd associations between contexts and words. A new class is generated by specification if we make the union of the constituent contexts (intension expansion) while the words are intersected (intension reduction). A new class is generated by abstraction if the local contexts are intersected (intension reduction), while we make the union of the constituent words (extension expansion). Intersecting words and local contexts in an accurate way allows us to generate tight clusters with prototypical constituents.

2 Related Work

Local syntactic contexts have been largely used to extract classes of semantically similar words. Yet, approaches differ in the way they define word similarity. Some of them assume that two words are similar if they co-occur with a number of identical local contexts [4, 6]. Semantic similarity is then computed by using the whole set of local contexts associated to each word. Unfortunately, the contexts of a word are usually very heterogeneous and multidimensional. They impose different selection restrictions and then select for different semantic facets or senses of a word. For instance, the noun *organisation* appears, at least, in two different types of contexts: those selecting for temporal events (*organisation* of the party, to finish the *organisation*, etc.) and those

* This work has been supported by Ministerio de Educación y Ciencia of Spain, within the project GaricoTerm, ref: BFF2003-02866.

requiring institutions (hired by the *organisation*, the president of the *organisation*, etc.). Given such a contextual diversity, this word can be semantically associated to a list of very heterogeneous nouns: *procedure*, *action*, *company*, *ministry*,.... This "absolute" view of semantic similarity leads to collapsing heterogeneous contextual information onto a single axis.

In order to induce semantically homogeneous lists of words, other approaches do not compare the semantic similarity between words, but between $< word, context >$ pairs and sets of those pairs. These sets are perceived as lexico-semantic classes or selection types [8, 9]. Given two vocabularies, W and LC, which represent respectively the set of words and the set of local contexts, a lexico-semantic class is defined as a pair $< LC', W' >$, where $LC' \subseteq LC$ and $W' \subseteq W$. In this model, the same word or context can in principle belong to more than one class. So, the positive side of these approaches is that they try to take into account polysemy. Some difficulties arise, however, in the process of class generation. Those aproaches propose a clustering algorithm in which each class is represented by the centroid distributions of all of its members. This is in conflict with the fact that many words and local contexts can significatively involve more than one semantic dimension. As a result, the clustering method appears to be too greedy since it overgenerates many wrong associations between words and local contexts.

To avoid this problem, a more recent approach tried to limit the information contained in the centroids by introducing a process of "clustering by committee" [7]. The centroid of a cluster is constructed by taking into account only a subset of the cluster members. This subset, called "committee", contains the more representative members (prototypes) of a class. So, the main and more difficult task of such an approach is to first identify a list of committees, i.e., a list of semantically homogeneous clusters. Committees represent basic semantic classes of similar words and are useful for word classification.

Other approaches also try to identify homongenous clusters representing basic semantic classes. The main difference with regard to the former method is that each basic cluster is constituted, not by similar words, but by a set of similar local contexts [2, 1, 3]. The method is focused on computing the semantic similarity between syntactic local contexts. Words are no more seen as objects to be clustered but as attributes of contexts. These are taken as the objects of the clustering process. As local contexts turn out to be less polysemic than words, it is assumed that searching for classes of homogeneous contexts is an easier task than to find tight classes of semantically related words. The main problem, however, is that the basic clusters of contexts identified in the first step tends to be very small and specific. The average size of a basic cluster is only two members. In order to generate larger classes, most of these approaches require a second step with a greedy clustering process. Unfortunately, this greedy clustering step tends to overgenerate many context-word associations.

The method proposed in this paper belongs to the last type of approach. Our main contribution is the use of very restrictive operations (specification and abstraction) in the process of building tight clusters. Thanks to these operations, we solve the overgeneration problem.

3 Assumptions

Following the model introduced by *Formal Concept Analysis* [5], lexico-semantic classes are defined as bi-dimensional objects: one dimension is the intension definition, i.e., a set of similar contexts with the same selection restrictions. The other one is its extension, i.e., the set of words appearing in such contexts and satisfying their semantic requirements. When the intension is very specific because it contains a large set of contexts, then the extension tends to be small.

New lexico-semantic classes are generated by means of a clustering process endowed with two complementary operations: specification and abstraction. If two similar classes, CL_1 and CL_2, defined respectively as the pairs $< LC_1, W_1 >$ and $< LC_2, W_2 >$, are aggregated into a new class, we can opt for two different operations:

specification: $CL_1 \Theta CL_2$, which represents a more specific class whose intension is the set of contexts $LC_1 \cup LC_2$, and the extension the word set $W_1 \cap W_2$.

abstraction: $CL_1 \Phi CL_2$, which represents a more generic class whose intension is the intersection $LC_1 \cap LC_2$, and the extension the union $W_1 \cup W_2$.

The clustering method we will describe in the following section makes use of these two operations. The resulting classes generated by such operations will be the startpoint of a further process: word classification.

4 The Method

Our method consists of 3 steps. In Step I, we describe a clustering algorithm relying on an specification operation. The aim is to extract a set of very specific classes. In Step II, these classes are merged by a hierarchical clustering and the abstraction operation. Finally, in Step III, each word is assigned to its more appropriate classes.

4.1 Step I: Extracting Specific Classes

We start by selecting a set of local syntactic contexts. As these contexts will be used as semantic word classifiers, they should not have high word dispersion. The word dispersion of a context is defined as the number of word types occurring with this context divided by the total number of word types in the training corpus. The input set is thus constituted by those contexts whose word dispersion is lower than an empirically set threshold.

Then, for each local context with low dispersion, we compute its top-*k* similar ones, where $k = 5$, using the weighted jaccard coefficient defined in [4] as a similarity measure. The extraction of specific classes operates on these ranked list. Given the top-5 list associated to a local context (and the set of word types it classifies), we first build 5 ranked classes by aggregating that context to each one in the list. Table 1 shows the five classes associated to the context "threat to [N]" that were extracted from the corpus *Europarl*. They will be the input of the clustering process.

The first class, 00231, is taken as the centroid since it is constituted by the top-1 similar context to "threat to [N]". The clustering process consists in aggregating the

Table 1. The top-5 classes built around the context "threat to [N]"

00231	{threat to [N], risk to [N]}	{health, environment, security, price, peace, stability}
00232	{threat to [N], endanger [N]}	{whole, democracy, peace, life, health, environment, security, stability}
00233	{threat to [N], [N] aspect}	{welfare, safety, employment, health, security}
00234	{threat to [N], damage [N]}	{employment, integrity, peace, life, health, environment, fishing, stability}
00235	{threat to [N], guarantee of [N]}	{safety, democracy, peace, job, freedom, security, stability}

remaining classes together around the identified centroid if only if they share more than 50% of the words. All aggregations are made using the operator of "specification" since each generated class is obtained by intersecting the two word sets of each aggregated class. As a result, we obtain:

CL_{37}	{endanger [N], damage [N], threat to [N], risk to [N]}	{health, environment, peace, stability}

which is the result of two specification operations:

$$CL_{37} = 00231 \; \Theta \; 00232 \; \Theta \; 00234$$

Here, clustering involves the centroid, 00231, and two classes, 00232 and 00234, which satisfy the similarity condition (share at least 50% of words). This process is repeated for all the ranked top-5 lists. The specific classes generated at the end of the process are the input of the following clustering step.

4.2 Step II: Generating Abstract Classes by Hierarchical Clustering

A standard hierarchical clustering takes the specific classes built in the previous step to generate new classes. For this purpose, we make use of an open source software: Cluster 3.0[1]. In this step, we use the operation of abstraction to build the successive aggregations. So, each generated class is constituted by both the union of word sets and the intersection of contexts. Table 2 illustrates a generic class, $NODE_{77}$, obtained from two successive abstractions: Table 2 illustrates a generic class, $NODE_{77}$, obtained from two successive abstractions:

$$NODE_{77} = CL_{37} \, \Phi \, NODE_{30}$$
$$NODE_{30} = CL_{420} \, \Phi \, CL_{202}$$

[1] *http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/software.htm*

Table 2. Hierarchical construction of the generic class $NODE_{77}$

$NODE_{77} : NODE_{30}\, \Phi\, CL_{37}$	{endanger [N]}	{health, life, patient, environment, peace, stability, quality}
$NODE_{30} : CL_{202}\, \Phi\, CL_{420}$	{endanger [N], risk to [N]}	{health, life, patient, environment, quality}
CL_{202}	{*endanger [N]*, risk to [N], expense of [N]}	{*health*, life, patient, environment}
CL_{420}	{*endanger [N]*, risk to [N], plant [N]}	{*health*, life, quality}
CL_{37}	{damage [N], *endanger [N]*, risk to [N], threat to [N]}	{*health*, environment, peace, stability}

Words and contexts organised around $NODE_{77}$ seem to characterise the abstract class of "entities in danger". Note that the classes we are able to learn (e.g., entities in danger) do not try to represent word senses as the synsets do in WordNet. Rather, they characterise ontological concepts.

The same word can appear in different generic classes. For instance, *environment*, which is a member of $NODE_{77}$, is also a member of another class aggregating nouns such as *agriculture*, *interior*, *justice*, *culture*, and *finance*, by their association with contexts like "minister of [N]", "ministry of [N]", or "minister for [N]".

Finally, if we observe more carefully Table 2, we find out that *health* and "*endanger [N]*" are the only elements appearing in the three specific classes. They can be considered as the prototypical or more representative constituents of such classes (they are in italic in the table). Prototypical elements will play an important role in the following step: word classification.

4.3 Step III: Word Classification

So far, the generated clusters have been loosing relevant information step by step, since they were aggregated using intersecting operations. Besides that, the intersecting aggregations did not allow us to infer context-word associations that were not attested in the training corpus. As has been mentioned above, our objective was to design a very restrictive clustering strategy so as to avoid overgeneralisations.

In order to both reintroduce lost information and learn new context-word associations, the last step aims at assigning more words to the specific classes generated in the first clustering process. A word is assigned to one or more classes in the following way:

We start by identifying the centroids used for classification. Given a specific class, the representative centroid is constituted by the words and contexts that were considered as prototypes in the abstraction process (Step 2). For instance, the centroid of prototypes extracted from the classes CL_{420}, CL_{202}, and CL_{37}, during the construction of $NODE_{77}$ is: $< \{endanger[N]\}, \{health\} >$. If a word fills the *classification conditions* imposed by this centroid, then it is assigned to the three classes in the example.

The classification conditions that a candidate word must fill are two: First, it must be *similar* to those words appearing in the centroid. Second, it must occur in the training corpus with the contexts of the centroid.

To measure similarity between words, we used the same coefficient as for context similarity: a weighted jaccard score. In addition, each word was provided with a list containing its top-5 most similar ones. So, two words, w_i an w_j, are considered to be similar if only if w_i is in the top-5 list of w_j, or conversely, if w_j is in the top-5 list of w_i.

At the end of the classification step, our system was able to assign "security", "democracy" , "growth", and "energy" to the classes organised around the concept of *entities in danger*. Note that the acquired classes refer to domain-dependent concepts.

5 Experiments and Evaluation

Experiments have been carried out using two different text corpora. A Portuguese corpus with 10 million tokens extracted from the general-purpose journal *O Público*, and an English excerpt (3 million tokens) of the European Parliament Proceedings (*EuroParl*), available in *http://people.csail.mit.edu/koehn/publications/europarl/*. Both corpora were *POS* tagged using TreeTagger[2], an open source software.

Table 3 depicts the number of specific and generic classes extracted from each corpus, as well as the number of word classifications. The extraction was only focused on nouns and nominal contexts. Note that not many generic classes were learnt. This is in accordance with the basic ideas underlying formal ontology.

Table 3. Corpus Data

	Specific Classes	**Generic Classes**	**Classifications**	**Accuracy of Classif.**
Público	264	91	492	92%
EuroParl	227	68	226	94%

Measuring the correctness of the acquired lexico-semantic classes is not an easy task. We are not provided with a gold standard against to which results can be compared. As the acquired classes are corpus-dependent and do not represent word senses, there is no pre-existing ontology nor thesaurus containing the type of information our system is able to learn. Some of the classes we learnt refer to particular encyclopaedic knowledge, for instance, the class of world regions with internal conflicts: *kosovo, balcans, serbia, colombia, chechnya, east timor, sierra leona, region*. These words appear in contexts such as "conflict in [N]" and "war in [N]". Encyclopaedic classes are, in general, semantically homogeneous and are organised in closed sets of words. By contrast, the system also acquires very heterogeneous classes constituted by open sets of words: e.g., entities in danger ($NODE_{77}$ above), different forces that can be involved in a process (*threat, obstacle, access, impetus, contribution, ...*), etc.

[2] *http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html*

To evaluate the lexico-semantic classes, we set a subjective evaluation protocol focused on the accuracy of word classification. Each word assignment to a class was judged as correct or incorrect by a human evaluator. An assignment was considered as correct if the assigned word is *semantically required* by all the local contexts defining the class. the 4th column of Table 3 shows the accuracy score. In fact, this evaluation measures the amount of overgeneration produced by the system. Overgeneration is about 8% in *O Público* and 6% in *EuroParl*.

In further research, we have to develop a process of context classification. In this process, each lexico-semantic class will be assigned local contexts that were not involved in the previous clustering step.

References

1. P. Allegrini, S. Montemagni, and V. Pirrelli. Example-based automatic induction of semantic classes through entropic scores. *Linguistica Computazionale*, pages 1–45, 2003.
2. 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, 2000.
3. Pablo Gamallo, Alexandre Agustini, and Gabriel Lopes. Clustering syntactic positions with similar semantic requirements. *Computational Linguistics*, 31(1):107–146, 2005.
4. Gregory Grefenstette. *Explorations in Automatic Thesaurus Discovery*. Kluwer Academic Publishers, USA, 1994.
5. Joachim Hereth, Gerd Stumme, Rudolf Wille, and Uta Wille. Conceptual knowledge discovery and data analysis. In *International Conference on Conceptual Structures*, pages 421–437. Berlin:Springer Verlag, 2000.
6. Dekang Lin. Automatic retrieval and clustering of similar words. In *COLING-ACL'98*, Montreal, 1998.
7. Patrick Pantel and Dekan Lin. Discovering word senses from text. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 613–619, Edmonton, Canada, 2002.
8. 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.
9. Mats Roth. Two-dimensional clusters in grammatical relations. In *Symposium on Representation and Acquisition of Lexical Knowledge: Polysemy, Ambiguity, and Generativity (AAAI 1995)*, 1995.