ADN-Classifier: Automatically Assigning Denotation Types to Nominalizations

Aina Peris*, Mariona Taulé*, Gemma Boleda, Horacio Rodríguez****
*CLiC, Centre de Llenguatge i Computació - University of Barcelona
Gran Via de les Corts Catalanes 585, 08007 Barcelona
{aina.peris, mtaule}@ub.edu
**TALP Research Center - Technical University of Catalonia
Jordi Girona Salgado 1-3, 08034 Barcelona
{gboleda, horacio}@lsi.upc.edu

Abstract

This paper presents the ADN-Classifier, an Automatic classification system of Spanish Deverbal Nominalizations aimed at identifying its semantic denotation (i.e. event, result, underspecified, or lexicalized). The classifier can be used for NLP tasks such as coreference resolution or paraphrase detection. To our knowledge, the ADN-Classifier is the first effort in acquisition of denotations for nominalizations using Machine Learning.We compare the results of the classifier when using a decreasing number of Knowledge Sources, namely (1) the complete nominal lexicon (AnCora-Nom) that includes sense distictions, (2) the nominal lexicon (AnCora-Nom) removing the sense-specific information, (3) nominalizations' context information obtained from a treebank corpus (AnCora-Es) and (4) the combination of the previous linguistic resources. In a realistic scenario, that is, without sense distinction, the best results achieved are those taking into account the information declared in the lexicon (89.40% accuracy). This shows that the lexicon contains crucial information (such as argument structure) that corpus-derived features cannot substitute for.

1. Introduction

This paper presents the ADN-Classifier, an Automatic classification system of Spanish Deverbal Nominalizations aimed at identifying its semantic denotation (i.e. event, result, underspecified, or lexicalized). The initial purpose of the ADN-classifier is to enrich the annotation of deverbal nominalizations in AnCora-Es corpus (Taulé et al., 2008) with this kind of information. Both the corpus and the machine learning experiments can give also more insight into the underlying linguistic question. Furthermore, the classifier can be used independently in other NLP tasks, such as coreference resolution or paraphrase detection. Regarding the first one, it is important to note that event and result nouns require different types of anaphoric pronouns in some languages. For instance, in Catalan the pronoun 'ho' refers to an event nominalization ('La matriculació de tots els alumnes es fa cada setembre. _Ho_ permet l'entorn informàtic de la universitat[1]) whereas 'el/la' refer to result nouns (eg. 'La construcció és molt innovadora. _L_'ha dissenyada el millor arquitecte del moment'[2]). Therefore, to have them classified can be useful to detect coreference chains. As for the second one, event nouns (but not result nouns) are paraphrases for full sentences, so this type of information can also be useful for paraphrase detection.

In Peris et al. (2009) a set of experiments were carried out in order to detect the most relevant features for the denotative distinction between event and result nominalizations. In these experiments, the foundations of the automatic classification system presented here were set. However, the experiments were mainly based on information manually coded in the lexicon. This approach does not easily extend to nominalizations and examples not covered in the lexicon. Here we evaluate to what extent this type of information can be recovered directly from the corpus AnCora-Es.

The paper is organized as follows. First, in section 2 we briefly discuss related work. Next, in section 3 we describe the ADN-Classifier, the different components and the resources used: AnCora-Es and AnCora-Nom. We then present in section 4 the results obtained. And finally, the main conclusions and final remarks are given in section 5.

2. Related Work

Related work on the computational treatment of nominalizations goes mainly in two directions. The first one is automatically classifying semantic relations in noun phrases. Task 4 of Sem-Eval 2007 shows a wide variety of automatic methods aiming to the classification of semantic relations (Girju et al., 2009).

More specific works related to nominalizations are Girju et al. (2004) and Lapata (2002). The former distinguish 35 semantic relations within NPs in which either the head or the modifier noun is derived from a verb. Some of these semantic relations can be seen as thematic role/argument relations (AGENT, THEME, TEMPORAL, CAUSE, EXPERIENCER, LOCATION, and PURPOSE). The latter focuses on the task of interpreting the semantic relations "subj" or "obj"[3] between the nominalization and its modifiers.

The second line of research consists of benefiting from verbal data to interpret, represent and assign semantic roles to nominalizations. Hull & Gomez (2000) design a series of algorithms that use verbal information (verbal senses and subcategorization frames) taking into account some specific noun constraints (specific order of noun arguments, constituents requirements for argument realization, different preposition in prepositional phrases

[1] 'The inscription of the pupils is every Setember. This is permitted by the univerty's software.'

[2] 'The building is very innovative. It has been designed by the most famous architect at the moment'.

[3] Lapata uses these terms for identifying Arg0 and Arg1 thematic roles relations, respectively.

complements, etc.) to interpret the nominalization NP. Gurevich et al. (2006) present a method for mapping deverbal nouns arguments to those of the corresponding verbs, relying on a rich lexicon and a series of heuristics, for knowledge representation purposes. In contrast, Padó et al. (2008) address the task of semantic role labeling for event nominalizations only from verbal training data, that is, they borrow existing verbal annotations to classify unseen but similar nominal predicates.

Most of these works are aware of the linguistic distinction between event and result nominalizations; however, none of them considers the distinction in their systems. We tackle precisely this task. Most closely related to our work are the efforts that are being done for German *'–ung'* nominalizations (Eberle et al., 2009), although their approach is symbolic and is focused only in nominalizations of verbs of information.

3. ADN-Classifier

We face the problem of classification examining the task performance when using a decreasing number of Knowledge Sources. First, following (Peris et al., 2009), we use as main Knowledge Source the complete nominal lexicon (AnCora-Nom) that includes sense distictions. We apply the model developed with this resource, to the initial set (100K words) from which the lexicon was built. In a second stage, we enrich the model with additional features extracted from AnCora-Es corpus. We apply the model to the same dataset (100K). We named these models sense-based.

Next, we reduce our dependence from the lexical source removing the sense-specific information while maintaining the features extracted from the corpus. We name these models lemma-based. In lemma-based models, when extracting features from the lexicon, we use as features for the classification those attributes whose values are shared by all senses of the same lemma. When no value is specified, the null value is assigned to the attributes.

Finally, we apply a model using only corpus derived features. For these two latter cases, we applied the models to the same dataset (100K). In this way, the setup of the task moves to a more realistic scenario.

3.1 Classification

We consider two basic semantic types of nouns, according to the linguistic literature (Grimshaw, 1990; Pustejovsky, 1995; Picallo, 1999):

- *Event* nouns: Those nouns denoting an action, in a similar way to verbs (e.g. 'La aprobación en febrero de este año de la supresión de 20 asientos de representación proporcional'[4,5]).
- *Result* nouns: Nouns that denote the result of an action (e.g. 'Contar con la aprobación del Consejo_de_Ministros'[6]).

Furthermore, in order to account for the data in the corpus we introduce two additional types:

- *Underspecified* nominalizations: In some cases, the linguistic context of the nouns does not allow us to disambiguate between the two denotations above (e.g. 'Anunció que el gabinete ha aprobado varias medidas económicas […]; la aprobación del proyecto de ley de telecomunicaciones, e incentivos a la inversión'[7]). We label those as underspecified.
- *Lexicalized* noun constructions: Cases in which the nominalization takes part in a lexicalized construction. In such cases, we distinguish among six types of lexicalizations according to their equivalence to different word classes: nominal ('síndrome de abstinencia'[8]), verbal ('estar de acuerdo'[9]), adjectival ('al alza'[10]), adverbial ('con cuidado'[11]), prepositional ('en busca de'[12]), or conjunctive ('en la medida que'[13]). Only in the case of nominal lexicalizations, one of the three previous denotative values (event, result, underspecified) is assigned. The remaining lexicalizations are assigned to the additional class *lexicalized.*

Thus, we model the task as a four way classification problem, with target classes event, result, underspecified, and lexicalized.

3.2 Method

Three stages are planned for the annotation of the corpus AnCora-Es and the evaluation of the ADN-Classifier: (1) evaluation with the nominalizations in the lexicon AnCora-Nom (817 lemmata) and a subset of the corpus (100K words; 3,077 examples manually labeled for the sake of the evaluation); (2) evaluation with the lemmata in the lexicon and the whole corpus (500K words); (3) evaluation with all the nominalizations in the corpus (1,662 lemmata, including the ones in (1)) and the whole corpus. Here we present the results of the first stage.

3.3 ADN-Classifier Architecture

In Figure 1, we present a schematic architecture of the ADN-Classifier.

The ADN-classifier consists of:

1) An extraction component, that uses the AnCora-Es corpus and the AnCora-Nom lexicon (Peris et al., 2009) to obtain the linguistic features for the learning and the classification processes.
2) A classifier component that uses the features extracted to classify the deverbal nominalizations into event, result, underspecified and lexicalized.

[4] All the examples are obtained from AnCora-Es corpus (http://clic.ub.edu/ancora/).

[5] 'The approval in February of this year of the suppression of 20 seats of proportional representation'.

[6] 'To rely on the approval of the Ministers' Council'.

[7] 'He announced that the cabinet has approved several economic measures […]; the approval of the project of law of telecommunications, and incentives to the investment'.

[8] 'Withdrawal symptoms'.

[9] 'To agree'.

[10] 'Upward'.

[11] 'Carefully'.

[12] 'Looking for'.

[13] 'As far as'.

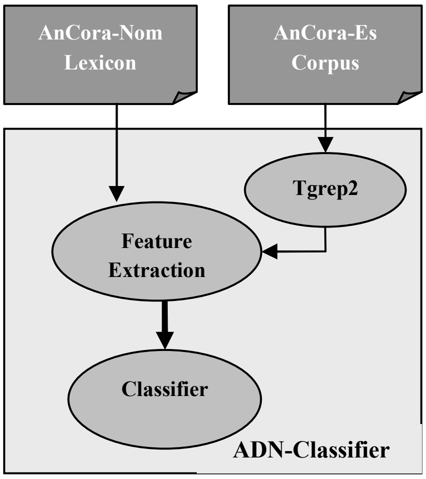


Figure 1: ADN-Classifier

3.3.1 Feature Extraction

To carry out the comparison between manually coded and corpus-derived information, two kinds of features are considered. Peris & Taulé (2009) present the most relevant features obtained from the analysis of the linguistic criteria from the literature (Grimshaw, 1990; Picallo, 1999; Alexiadou, 2001) to establish the distinction between both denotations. The selected features are the following:

Features from the AnCora-Nom lexicon. These include information on the semantic class of the verb from which the deverbal noun is derived (as specified in the verbal lexicon AnCora-Verb-Es; Aparicio et al., 2008 [14]), information on complementation and argument structure, and other features related to the NPs the nominalization appears in: type of specifier (definite, indefinite, possessive, demonstrative, etc.); number (plural or singular); whether the nominalization takes part in a lexicalized construction; suffix[15]. Note that the lexicon contains information both about the lemma and about the examples that will be used in the experiment (see Figure 2).

In Figure 2, the sense 2 of *aumento*[16] lexical entry is shown. The lexical features represented are:

a) The semantic class of the verb from which the nominalization is derived. In the example, the transitive frame of *aumentar_1* (originlink="verb.aumentar.1.transitive") corresponding to the semantic class a2 (accomplishment agentive-transitive). The atributte "originlink" is also used to relate the nominal sense to the verbal one. In this way, the nominal and the verbal lexicons are linked.

b) The nominal complements ("constituent type" attribute) and the argument structure with the corresponding theta roles ("argument" and "thematicrole" features, respectively). The sense 2 of *aumento* can take two arguments: an Arg1 with the thematic role "tem" (*theme*) that can be realized by a PP ("sp") introduced either by the preposition "de" or "en"; and an Arg2 bearing the thematc role "ext" (*extension*) which can be realized by a PP introduced by the preposition "de". The third possible complement is a non-argumental adjective phrase ("s.a").

c) The type of specifiers that the noun sense appears with in the corpus. The example shows that *aumento_2* occurs without any specifier ("void") or with an indefinite article ("indef").

d) The number in which that particular noun sense appears in the corpus. Since at least one of the examples of the *aumento_2* is in plural, the value of the feature "plural" is positive (frame canbeplural="yes").

e) Whether the nominalization takes part in lexicalized construction or not and which type of nominalization it is. (Figure 2 does not contain a lexicalized nominalization sense, therefore this feature is not shown).

```
<lexentry lemma="aumento" lng="es" origin="deverbal">
<sense id="2">
   <frame canbeplural="yes"
originlink="verb.aumentar.1.transitive"
type="result">
      <argument argument="arg1" thematicrole="tem">
          <constituent preposition="de" type="sp"/>
           <constituent preposition="en" type="sp"/>
      </argument>
      <argument argument="arg2" thematicrole="ext">
         <constituent preposition="de" type="sp"/>
      </argument>
      <nonargumental>
         <constituent type="s.a"/>
      </nonargumental>
      <specifiers>
          <constituent type="void"/>
         <constituent type="indef"/>
      </specifiers>
      <examples>
        <example>
Sometidos a un dopado oxidante muestran aumentos
sustanciales de la conductividad .
</example>
<example>
Las exportaciones registraron un aumento del  3,3_por_ciento
 </example>
      </examples>
    </frame>
</sense>
```

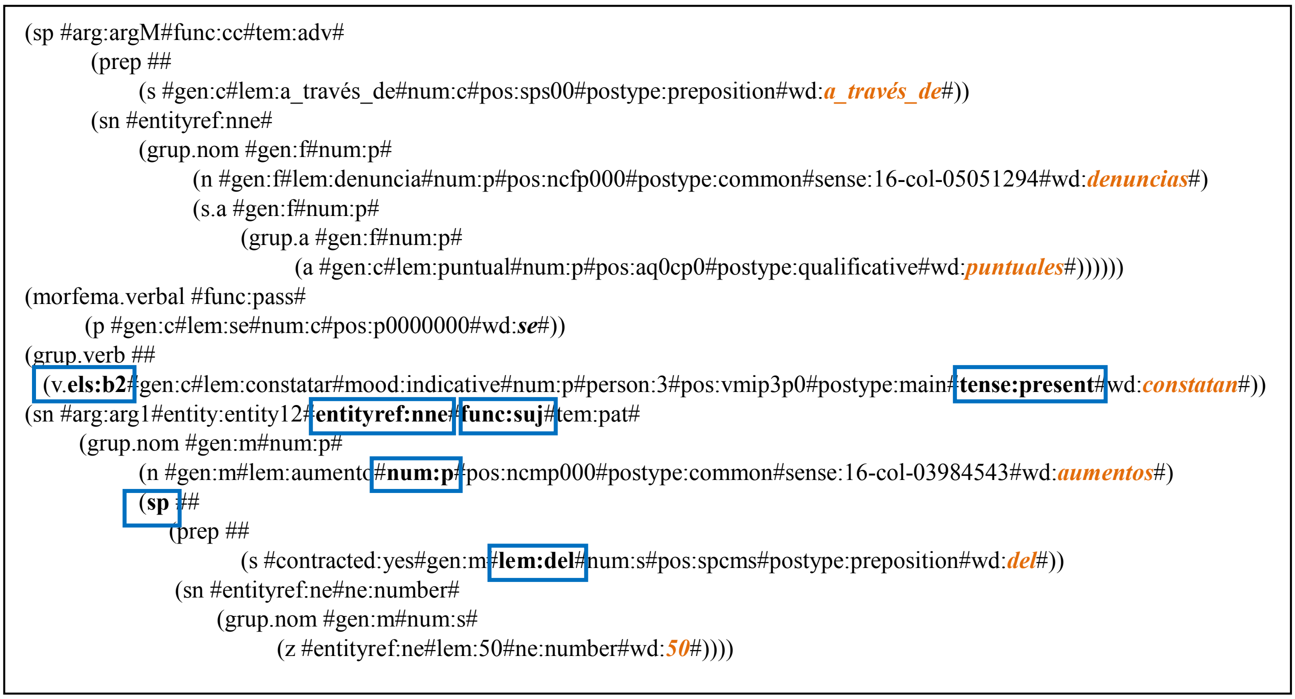
Figure 2: Partial lexical entry of *aumento* [17]

[14] In AnCora-Verb-Es lexicon each predicate is related to one or more semantic classes (Lexical Semantic Structure, LSS) depending on its senses, basically differentiated according to the four event classes ─accomplishments (A), achievements (B), states (C) and activities (D)─, and on the diatheses alternations in which a verb can occur.

[15] This attribute is taken from a predefined list of suffixes: *-ción*, *-aje*, *-ido*, *-do*, *-da*, *-ura*, *-encia*, *-enza*, *miento*, *-mento*, *-o*, *-a*, *-e*.

[16] 'Increase'.

[17] This figure shows only the sense 2 of the *Aumento* lexical entry. The whole lexical entry is available in http://clic.ub.edu/ancora/.



```
(sp #arg:argM#func:cc#tem:adv#
        (prep ##
              (s #gen:c#lem:a_través_de#num:c#pos:sps00#postype:preposition#wd:a_través_de#))
        (sn #entityref:nne#
              (grup.nom #gen:f#num:p#
                     (n #gen:f#lem:denuncia#num:p#pos:ncfp000#postype:common#sense:16-col-05051294#wd:denuncias#)
                     (s.a #gen:f#num:p#
                           (grup.a #gen:f#num:p#
                                  (a #gen:c#lem:puntual#num:p#pos:aq0cp0#postype:qualificative#wd:puntuales#))))))
(morfema.verbal #func:pass#
        (p #gen:c#lem:se#num:c#pos:p0000000#wd:se#))
(grup.verb ##
  (v.els:b2#gen:c#lem:constatar#mood:indicative#num:p#person:3#pos:vmip3p0#postype:main#tense:present#wd:constatan#))
(sn #arg:arg1#entity:entity12#entityref:nne#func:suj#tem:pat#
       (grup.nom #gen:m#num:p#
              (n #gen:m#lem:aumento#num:p#pos:ncmp000#postype:common#sense:16-col-03984543#wd:aumentos#)
              (sp ##
                  (prep ##
                           (s #contracted:yes#gen:m#lem:del#num:s#pos:spcms#postype:preposition#wd:del#))
                  (sn #entityref:ne#ne:number#
                        (grup.nom #gen:m#num:s#
                               (z #entityref:ne#lem:50#ne:number#wd:50#))))
```

Figure 3: A partial parse tree from AnCora of the sentence
"a_través_de denuncias puntuales se constatan aumentos del 50%"
('Through punctual claims increases of 50 % are observed')

When building the lexicon, the different readings a particular lemma shows in the corpus (say, *event* and *result*) are identified and the examples corresponding to each reading are grouped in the lexical entry. In Figure 2, the *aumento* sense 2 that corresponds to a result reading is associated to two examples. Therefore, the properties found in any particular example are shared by the remaining examples of the relevant sense of the lemma.

Since we have 817 lemmas and 3,077 examples, the average number of examples per lemma is 3.8.

Also, note that we have experimented with binarization and grouping of several of the features. Due to the excessive dispersion of the values in some features, we have grouped some of the values in order to facilitate the learning process. For example, in the *sp* feature, without grouping, the number of possible values is 101, which is too high. Two types of groupings have been considered: i) one that takes into account the number of the argument (Arg0, Arg1, Arg2, Arg3, Arg4, ArgM in addition to the no argument value, CN, thus, 7 possible values); and ii) a fine grained one, which incorporates to the code the preposition involved (Arg0-con, Arg0-de, etc.), giving rise to 60 possible values (for details, see Peris et al. 2009).

Features from the AnCora corpus. These features contain information that can be extracted directly from the syntactic tree of the corpus. Any information present in the treebank but not directly derived from the parse tree is not taken into account.

In Figure 3, a partial parse tree with all the associated morphosyntactic and semantic information is presented. From these trees we obtained the following features:

a) The corpus versions of the features from the lexicon specified above: the type of specifier, the number (plural or singular), the constituent type of the complements.
b) Other contextual features such as tense and semantic class of the main verb in the sentence; syntactic function of the nominalization; and whether the noun appears in a named entity.

For the extraction of features from the corpus, we use the Tgrep2[18] tool, which allows us to efficiently inspect the syntactic trees in a Treebank format. Below, we show an example of a tgrep2 rule that illustrates the extraction of the feature *plural*. The feature *plural* is very useful since is taken by the classifier as a rule for detecting result nominalizations:

```
tgrep2 -c AnCora_Es.tbf.t2c -a -l -w '(sn < ("grup.nom" <
(n < /aumento/ < /num:p/)))'
```

This pattern can be paraphrased as "look for a NP ("sn") that dominates immediately a nominal group ("grup.nom"), which, in turn, dominates a noun ("n") with the lemma *aumento* that appears in plural ("num:p").

In the treebank shown in Figure 3, the tgrep2 rule above will set the *plural* feature to "True" since *aumento* appears in plural in the example, which is marked in the *number* attribute (cf. "num:p").

Other tgrep2 rules will allow us to recognize the other corpus features. In Figure 3, the tgrep2 rules detect that in this example, the nominalization is not specified, that is, it does not have a determiner; the nominalization complement is a prepositional phrase introduced by the preposition "de"; the main verb in the sentence

[18] http://tedlab.mit.edu/~dr/TGrep2/

("constatan"[19]) occurs in present ("tense:present") and has a "b2" semantic class ("els:b2")[20]; the syntactic function of the nominalization is subject ("func:suj"); and it does not appear in a named entity ("entityref:nne").

Next, we present a table which summarizes the number and types of features used in the lemma-based (lexicon + corpus) model of the ADN-classifier (see Section 4).

FEATURE TYPES		NUMBER
AnCora-Nom	Simple	23
	Binarized	169
AnCora-Es	Corpus lexicon version	35
	Basic contextual	44
	Derived contextual	13
TOTAL		**284**

Table 1: Number and types of features used in ADN-classifier.

The features from AnCora-Nom lexicon include:

a) A total of 14 simple features (noun_type, semantic verbal class, number, suffix, specifier, possessive specifier, PP, AP, NP, ADVP, relative, sentence and lexia) with their corresponding groupings (9). The total amount of simple features is 23.

b) A total of 169 binarized versions of such features (for details, see Peris et al. 2009).

The features extracted from AnCora-Es corpus are:

a) The corpus versions of the features from the lexicon related to number, specifier, type of complement and preposition introducing PPs. The features involved are 35.

c) A total of 44 simple contextual features such as tense and semantic class of the main verb in the sentence; syntactic function of the nominalization; and, whether the noun appears in a named entity, which are extracted with a unique tgrep2 rule.

d) A total of 13 complex contextual features resulting from the combination of two or more tgrep2 rules. One example is the feature that takes into account whether the nominalization acts as subject or direct complement of a duration verb (tipically denoting an *event* noun). We combine the tgrep2 rule that obtains the syntactic function of the nominalization and another that extracts the verb lemma and check whether it is in a predefined list of verb duration lemmas[21].

3.3.2 Classifier

For the present experiments we use a rule-based classifier (J48.Part, the rule version of the tree decision classifier C4.5; see Quinlan, 1993) as implemented in the software Weka (Witten & Frank, 2005). We have chosen a decision tree classifier because it provides a natural representation of classification rules, thus allowing for inspection of the model without drop in accuracy. We have used the Weka framework both for learning and classifying.

4. Results

Recall that the task is to classify the nominalizations in the 3,077 sentences as (a) *result*, (b) *event*, (c) *underspecified*, or (d) *lexicalized*.

Table 2 presents the overall results. The columns contain the models used (see Section 2), the number of features in each model[22], the number of rules built by the classifier and the accuracy obtained. Recall is always 100%, so accuracy and precision are the same.

The rows correspond to different models presented in a decreasing order of Knowledge Sources as described in Section 2. The last row shows the baseline, a majority baseline which assigns all examples to class *result*. Note that the baseline is very high: there are over 80% result nominalizations in the corpus.

Model	Features	Rules	Accuracy
Sense-based model (Lexicon only)	251	61	93.6
Sense-based model (Lexicon +corpus)	359	99	92.8
Lemma-based model (Lexicon only)	193	92	89.4
Lemma-based model (Lexicon + corpus)	284	148	87.5
Lemma-based model (Corpus only)	97	109	80.6
Baseline	0	1	82.1

Table 2: Results of the classification experiments with different types of features.

As can be seen in Table 2, the sense-based models outperform the lemma-based models. These represent the upper bound for our task. However, in a realistic scenario, given the state of the art results in Word Sense

[19] 'To state'.

[20] The els b2 here corresponds to the passive voice of the agentive-transitive (a2) default semantic class of *constatar*.

[21] The predefined list of duration is below: *preparar* ('to prepare'), *arreglar* ('to fix), *empezar* ('to begin), *comenzar* ('to start'), *iniciar* ('to initiate), *entrar* ('to enter), *lanzar* ('to throw), *reiniciar* ('to reinitiate), *reempezar* ('to re-begin'), *recomenzar* ('to restart'), *nacer* ('to born), *durar* ('to last), *seguir* ('to follow'), *continuar* ('to continue'), *mantener* ('to maintain'), *progresar* ('to progress), *congelar* ('froze'), *parar* ('to stop'), *interrumpir* ('to interrupt'), *suspender* ('to suspend), *acabar* ('to finish'), *completar* ('to complete), *concluir* ('to conclude), *terminar* ('to end'), *culminar* ('to culminate'), *abandonar* ('to leave'), *cesar* ('to stop'), *finalizar* ('to finalize').

[22] Note that the number of features in the Lemma-based model (lexicon + corpus) does not correspond to the sum of the number of features in Lemma-based model (lexicon only) and Lemma-based model (corpus only). This is due to the fact that there are 6 features shared by both models and are taken into account only once in the combined model (lexicon + corpus). These 6 features correspond to some of the corpus versions of the lexicon features, as described in Section 3.3.1.

Disambiguation, we would not have access to sense labels, so we focus on the lemma-based models.

Without sense distinction, we achieve 89.4% accuracy, compared to an 82.1% baseline. As shown in the fifth row of Table 2, the corpus features yield accuracy values that are below the baseline. In fact, corpus-based features harm accuracy both in the sense and lemma based models. Thus, the information in the corpus, at least as currently coded, is not able to distinguish between the different readings of the nominalizations. From these results, it can be inferred that there is crucial information in the lexicon that is not possible to recover from the corpus, concretely the argument structure with the corresponding thematic roles of nominalizations complements. Furthermore, corpus based features suffer from data sparseness. As has been explained in Section 3.3.1, when using features from the lexicon the information found for one example is generalized to the remaining examples (from one example to 3.8 in average). However, in the corpus version each feature can only be associated to one particular example; therefore, the features are sparser.

We plan to use the lemma-based model that uses corpus information as well as the lexicon (fourth row in Table 2) for phase 2 of the corpus annotation (that is, the annotation of 500K examples with the same lemmata as in phase 1; see Section 3.2 above). Even if this model has a lower accuracy on the subset of the corpus considered so far (87.5% as opposed to 89.4%), we expect it to exhibit a more robust behaviour when tackling unseen data.

4.1 Analysis of Errors

The analysis of errors focuses on the lemma-based model with lexicon and corpus information, for the reasons outlined in the previous section. Table 3 shows the performance of this model for each class. As can be seen in the table, the model is very successful at detecting result and lexicalized nominalizations (91.9% and 95.4% F-Measure, respectively), but fails for the event and underspecified classes (54.6% and 56.4% F-Measure, respectively). The latter is to be expected, given that these are the cases with no clear contextual hints as to their class, or true ambiguous cases. The most serious problem for this model, and indeed for the whole enterprise, is the failure to detect event uses of nominalizations.

	precision	recall	F-Measure
R	90.7	93.1	91.9
E	56.9	52.5	54.6
L	95.0	95.8	95.4
U	63.9	50.5	56.4

Table 3: Class-based performance of the Lemma-based model (Lexicon + corpus). Legend: R = result, E = event, L = lexicalized, U = underspecified.

Table 4 presents the confusion matrix of the model. Rows correspond to manually labeled data and columns are predictions from the classifier. The correct predictions are in the diagonal (in bold face). The main sources of errors are marked in italics.

As can be seen in Table 4, most of the errors involve the distinction between event and result nominalizations: 109 event nominalizations are classified as result, and 79 result nominalizations are classified as event. As mentioned above, this is the biggest challenge for our model, and it clearly points to the fact that the features we are using do not contain enough information to successfully detect event nominalizations.

Correct ↓	R	E	L	U	Total
R	**1881**	*79*	22	*38*	2020
E	*109*	**145**	8	14	276
L	18	7	**574**	0	599
U	*66*	24	0	**92**	182
Total	2074	255	604	144	3077

Table 4: Confusion Matrix for the Lemma-based model (Lexicon + corpus). Legend as in Table 3.

We believe that the most plausible explanation for these results is the fact that the most useful information to detect event uses of nominalization is the presence of argument structure. Information on argument structure is sense-dependent, and it is hard or impossible to recover this information from the lexicon at a lemma level.

There are surface cues for the presence of an argument, such as a PP complements or possessive determiners. Thus, this information could in principle be recovered from the parse tree of each particular example (corpus features). However, the problem is that non-argument complements are syntactically realized by exactly the same type of constituents, so corpus features are often misleading.

The second main type of error is the misclassification of underspecified nominalizations: as shown in Table 4, 66 underspecified cases are classified as result, and 38 result nominalizations are classified as underspecified. In this case, we believe that the main problem is data sparseness, in the sense that, as mentioned above, it is not possible to determine the correct class only from the context.

5. Conclusions

Our goal is to build a general purpose classifier of Spanish deverbal nominalizations that uses information that could be extracted automatically, such as the morphosyntactic structure of the sentences. So far, the results with corpus based features are below the baseline. The versions of the classifier that consider the information in the lexicon achieve a 7% improvement over the baseline. This leads us to the conclusion that there is crucial information in the lexicon that is not possible to recover from the parse tree, such as the argument structure and the corresponding thematic roles of the nominalizations. We believe this will help us to detect more succesfully event nominalizations, which is the biggest challenge for our task. In fact, we are currently working on the automatic annotation of this information benefiting from the verbal data contained in AnCora-Verb.

Future work will focus on how to automatically obtain other types of information that help in the detection of event nominalizations, such as some aspects of discourse structure or semantic information contained in general purpose resources (e.g., dictionaries and WordNet).

Another line of improvement will consist in relaxing the constraint on the full coverage using a threshold to remove the least accurate rules. This procedure may be useful for the automatic annotation of AnCora corpus.

6. Acknowledgements

This research has received support from the projects Text-Knowledge 2.0 (TIN2009-13391-C04-04), AnCora-Nom (FFI2008-02691-E), AnCora-Net (FFI2009-06497-E/FILO), KNOW2 (TIN2009-14715-C04-04) and OntoSem (Hum2007-60599/FILO), from the Spanish Ministry of Science and Innovation, and the program Juan de la Cierva (JCI-2007-57-1479) and a FPU grant (AP2007-01028) from the Spanish Ministry of Education. We also thank three anonymous reviewers for their comments and suggestions.

7. References

Alexiadou, Artemis (2001). *The Functional Structure in Nominals. Nominalization and Ergativity*. Amsterdam / Philadelphia. John Benjamins.

Aparicio Aparicio, Juan., Mariona Taulé & M.Antònia. Martí (2008). AnCora-Verb: A Lexical Resource for the Semantic Annotation of Corpora. In *Proceedings of Language, Resources and Evaluation*, LREC 2008. Marrakech, Morocco.

Eberle, Kurt, Gertrud Faaß & Ulrich Heid (2009). Corpus-based identification and disambiguation of reading indicators for German nominalisations. *Corpus Linguistics 2009*. Liverpool, UK.

Grimshaw, Jane (1990). *Argument Structure*. Cambridge, Massachussets: The MIT Press.

Girju, Roxana, Ana Maria Giuglea, Marian Olteanu, Ovidu Fortu, Orest Bolohan & Dan Moldovan (2004). Support Vector Machines applied to the Classification of Semantic Relations in Nominalized Noun Phrases. In *Proceedings of HLT-NAACL. Computational Semanticc Workshop.* Boston, MA.

Girju, Roxana, Preslav Nakov, Vivi Nastase, Stan Szpakowicz, Peter Turney, & Deniz Yuret (2007). SemEval-2007 task 04: classification of semantic relations between nominals. In *SemEval '07: Proceedings of the 4th International Workshop on Semantic Evaluations*. Association for Computational Linguistics. Prague, Czech Republic.

Gurevich, Olga, Richard Crouch, Tracy Holloway King & Valeria De Paiva (2006). Deverbal Nouns in Knowledge Representation. In *Proceedings of FLAIRS.* Melbourne Beach, FL.

Hull, Richard D. & Fernando Gomez (2000). Semantic interpretation of deverbal nominalizations. In *Natural Language Engineering.* Cambridge University Press.

Lapata, Maria. (2002). The Disambiguation of Nominalizations. *Computational Linguistics.* Vol 23 (3). MIT Press.

Padó, Sebastian, Marco Pennacchiotti & Caroline Sporleder. (2008). Semantic role assignment for event nominalizations by leveraging verbal data. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling).* Manchester.

Peris, Aina & Mariona Taulé (2009). Evaluación de los criterios lingüísticos para la distinción evento y resultado en los sustantivos deverbales. In *Proceedings of the 1st International Conference on Corpus Linguistics* (CILC-09), Murcia, Spain.

Peris, Aina, Mariona Taulé & Horacio Rodríguez (2009). Hacia un sistema de clasificación automática de sustantivos deverbales. *Procesamiento del Lenguaje Natural*, 43. Jaén, Spain.

Picallo, Carme (1999). La estructura del sintagma nominal: las nominalizaciones y otros sustantivos con complementos argumentales. In Bosque & Demonte (Eds.) *Gramática Descriptiva de la Lengua Española*. Real Academia Española. Espasa Calpe, Madrid, Spain.

Pustejovsky, James (1995). *The Generative Lexicon. Cambridge*. MIT Press.

Quinlan, J. Ross (1993). *C4.5: Programs for Machine Learning*. San Francisco. Morgan Kaufmann.

Taulé, Mariona., M.Antònia. Martí & Marta Recasens (2008). AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of Language, Resources and Evaluation*, LREC 2008. Marrakech, Morocco.

Witten, Ian H. & Eibe Frank (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Franscisco, second edition.