IARG-AnCora: Annotating AnCora corpus with implicit arguments

Aina Peris, Mariona Taulé
CLiC-Centre de Llenguatge i Computació
University of Barcelona
Gran Via 585, 08007 Barcelona
{aina.peris,mtaule}@ub.edu

Horacio Rodríguez
TALP Research Center
Technical University of Catalonia
Jordi Girona Salgado 1-3, 08034 Barcelona
horacio@lsi.upc.edu

Abstract

IARG-AnCora is an ongoing project whose aim is to annotate the implicit arguments of deverbal nominalizations in the AnCora corpus. This corpus will be the basis for systems of automatic Semantic Role Labeling based on Machine Learning techniques. Semantic Role Labelers are essential components of current language technology applications in which it is important to obtain a deeper understanding of the text in order to make inferences at the highest level in order and thereby obtain qualitative improvements in the results.

1 Introduction

Traditionally, the analysis of argument structure has been focused basically on verbal predicates, although it has recently been extended to nominal predicates. Most of the efforts at argument identification are restricted to those arguments that appear in the sentence, in the case of verbs, or in the Noun Phrase (NP), in the case of nouns. In a nutshell, they are focused on the identification of explicit arguments. Furthermore, Semantic Role Labeling (SRL) systems are verb-centered and reduce role labeling to explicit arguments (Márquez et al., 2008; Palmer et al., 2010). In order to move forward to the full comprehension of texts, it is necessary to take into account implicit arguments and to widen the context of analysis to the whole discourse (Gerber et al., 2009). This is especially important in the case of deverbal nominalizations since the degree of optionality of their explicit arguments is higher than for verbs.

The aim of *IARG-AnCora* is to enrich the Spanish and Catalan AnCora corpora[1] by annotating the implicit arguments of deverbal nominalizations. Currently, AnCora corpora are only annotated with arguments inside the NP of these deverbal nouns. AnCora consists of a Catalan (AnCora-Ca) and Spanish (AnCora-Es) corpora of 500,000 words each, annotated at different linguistic levels: morphology (Part of Speech, PoS, and lemmas), syntax (constituents and functions), semantics (verbal and deverbal nouns argument structure, named entities and WordNet senses), and pragmatics (coreference). The main goal is to identify implicit arguments and assign an argument position –iarg0[2], iarg1, etc.– and a thematic role (agent, patient, cause, etc.) to them. These arguments can be recovered if a wider discursive context is taken into account and their identification is therefore important to provide a deep semantic representation of sentences and texts.

2 Defining an Implicit Argument

We define an implicit argument as the argument which is not realized in the NP headed by the deverbal nominalization, but is realized instead inside (1) or outside the sentence (2) context[3]. However, the implicit argument can sometimes be inside the NP

[1] AnCora corpora are freely available at: http://clic.ub.edu/corpus/ancora.

[2] The letter 'i' at the beginning of the argument position stands for implicit argument. We note the implicit arguments as iarg<position>-<thematic role>.

[3] We focus our definition of implicit arguments on deverbal nominalizations because we deal with them in our work. However, it is worth saying that verbs can also have implicit arguments.

as long as the constituent associated to this implicit argument does not depend directly on the nominalization. For instance, constituents inside a subordinate clause complementing the deverbal noun can be implicit arguments (3) of this deverbal noun.[4]

(1) [Las escuelas de samba de Sao Paulo]$_{\text{iarg1-pat}}$ han conseguido [el **apoyo**[5] [de la empresa privada]$_{\text{arg0-agt}}$ para mejorar las fiestas de carnaval]$_{\text{NP}}$.
*[Schools of samba in Sao Paulo]$_{\text{iarg1-pat}}$ got [the **support** [of private industry]$_{\text{arg0-agt}}$ to improve Carnival celebrations]$_{\text{NP}}$.*

(2) [El carnaval de Sao Paulo es feo]$_{\text{iarg1-pat}}$, dijo hoy [el alcalde de Río de Janeiro]$_{\text{iarg0-agt}}$ en una conversación informal con periodistas cariocas, y encendió la polémica. [...] [Esa **opinión**[6]]$_{\text{NP}}$ fue respaldada por el gobernador de Río de Janeiro, quien incluso fue más allá en su crítica al comentar que el carnaval que se organiza en Sao Paulo es "más aburrido que un desfile militar".
*[The Carnival of Sao Paulo is ugly]$_{\text{iarg1-pat}}$, said [the mayor of Rio de Janeiro]$_{\text{iarg0-agt}}$ in an informal conversation with Carioca journalists, and ignited the controversy. [...] [This **opinion**]$_{\text{NP}}$ was supported by the governor of Rio de Janeiro, who went even further in his criticism when he commented that the carnival held in Sao Paulo is"more boring than a military parade".*

(3) [El **daño** [causado a [su industria aeronáutica]$_{\text{iarg1-tem}^7}$]$_{\text{Subordinate C}}$]$_{\text{NP}}$.
*[The **damage** [caused to [its aeronautic industry]$_{\text{iarg1-tem}}$]$_{\text{Subordinate C}}$]$_{\text{NP}}$.*

Example (1) shows the deverbal nominalization 'apoyo' *support* with the agent argument ('de la empresa privada', *of private industry*) realized inside the NP, whereas the patient argument ('las escuelas de samba de Sao Paulo', *schools of samba in Sao Paulo*) is realized in the same sentence but outside the NP. In (2), the nominalization 'opinión', *opinion*, appears without any explicit argument in the NP. However, the agent argument ('el alcalde de Río de Janeiro', *the mayor of Rio de Janeiro*) as well as the patient argument ('el carnaval de Sao Paulo es feo', *the carnival of Sao Paulo is ugly*) are realized implicitly (iarg0-agt and iarg1-pat, respectively) in the previous sentence. Currently, the AnCora corpus is only annotated with arguments inside the NP, therefore 'opinión' *opinion* has no associated argument and 'apoyo' *support* only has the *agent* argument annotated. In example (3), the implicit argument of 'daño' *damage*, iarg1-tem, is the 'industria aeronáutica' (*aeronautic industry*), which is a constituent inside the subordinate clause.

3 Corpora annotated with implicit arguments

As far as we know, the only two corpora with nominal implicit arguments have been developed for English and they have been used as training data for the works presented in (Ruppenhofer et al., 2010) and (Gerber and Chai, 2010):

- The training and test corpus developed for SemEval-2010 task 10[8], *Linking events and their participants in discourse* (Ruppenhofer et al., 2010). A corpus that consists of literary texts annotated following FrameNet-style.

- A subset of the standard training, development, and testing sections of the Penn TreeBank (Marcus et al., 1993) used in (Gerber and Chai, 2010). The annotation scheme follows PropBank (Palmer et al., 2005) and NomBank (Meyers et al., 2004; Meyers, 2007) proposals.

The number of occurrences annotated is 3,073 in the former, where each nominal predicate had a very small number of occurrences, and 1,253 in the latter,

[4] In NomBank, these cases are annotated as arguments outside the domain of locality, and are therefore not treated as implicit arguments (Meyers, 2007). We only consider explicit arguments to be those that depend directly on the nominal predicate.

[5] In AnCora corpus, 'conseguir apoyo' is not considered to be a support verb construction because the verb is not semantically bleached and it holds a predicating power (Hwang et al., 2010), so 'apoyo' is annotated as the object of 'conseguir' and they are treated as independent predicates.

[6] In Spanish, the noun 'opinión', *opinion*, is derived from the verb 'opinar', *to express an opinion*.

[7] The label 'tem' stands for theme.

[8] `http://www.coli.uni-saarland.de/projects/semeval2010_FG/`.

where only the ten most frequent unambiguous noun occurrences are annotated in order to avoid the problem of sparseness presented in the SemEval-2012 task 10 corpus. Both corpora are annotated only with core arguments (no adjuncts arguments).

IARG-AnCora will be the first corpus annotated with implicit arguments in Spanish and Catalan. In contrast to the English corpora, IARG-AnCora will have an extended coverage in two senses: on the one hand, all the implicit arguments of all deverbal nominalization occurrences in the corpus AnCora (approximately 19,000 for each language) will be annotated; on the other hand, we will take into account the core arguments (arg0, arg1, arg2, arg3 and arg4) as well as the adjunct arguments (argM).

4 Methodology

We will annotate the implicit arguments of AnCora in three steps combining automatic and manual processes.We have already completed the first step and now we are focused on the second.

(a) First, we have developed a manually annotated training corpus consisting of 2,953 deverbal noun occurrences in AnCora-Es. These occurrences correspond to the 883 unambiguous deverbal nominalization lemmas, that is, to those that have only one sense (with only one roleset associated) in AnCora-Nom (Peris and Taulé, 2011a). In order to ensure the quality and the consistency of the annotated data, an inter-annotator agreement test has been conducted on a subsample of 200 occurrences. The average pairwise result obtained between the three pairs of annotators was 81% of observed agreement (58.3% Fleiss kappa (Fleiss, 1981)). The features for the classification model will be inferred from this training corpus.

(b) Second, we will develop an implicit argument SRL model based on Machine Learning (ML) techniques, whose purpose is the automatic identification and classification of implicit arguments. We will use this model to automatically annotate the implicit arguments of the whole AnCora-Es. Afterwards, we will adapt this model and apply it to Catalan (AnCora-Ca) in order to analyze its transportability[9].

(c) Finally, a manual validation of the automatically annotated corpus will be carried out in order to ensure the quality of the final resource. This manual validation will allow for the evaluation of the precision and recall of the automatic system developed.

In the automatic and the manual processes, we use the verbal and nominal lexicons -AnCora-Verb (Aparicio et al., 2008) and AnCora-Nom- as lexical resources to obtain the information about the possible implicit arguments for each predicate. The candidate arguments to be localized in the local discursive context, and to be thereafter annotated, are those specified in the nominal or verbal lexical entries and not realized explicitly.

4.1 Annotation Scheme

We use the same annotation scheme as the one followed to annotate the explicit arguments of deverbal nouns (Peris and Taulé, 2011b), and the argument structure of verbs in AnCora (Taulé et al., 2008), which was in turn based on PropBank and NomBank. In this way, we ensure the consistency of the annotation of arguments of different predicates -nouns and verbs-, as well as the compatibility of Spanish and Catalan resources with English resources.

We use the iarg$_n$ tag to identify implicit arguments and to differentiate them from explicit arguments (arg$_n$) (Gerber and Chai, 2010). The list of thematic roles includes 20 different labels based on VerbNet (Kipper, 2005) proposals: agt (agent), cau (cause), exp (experiencer), scr (source), pat (patient), tem (theme), cot (cotheme), atr (attribute), ben (beneficiary), ext (extension), ins (instrument), loc (locative), tmp (time), mnr (manner), ori (origin), des (goal), fin (purpose), ein (initial state), efi (final state), and adv (adverbial).

The combination of the six argument positions labels (iarg0, iarg1, iarg2, iarg3, iarg4, iargM) with the different thematic roles results in a total of 36 possible semantic tags (iarg0-cau, iarg1-agt, iarg0-agt, iarg2-loc, etc.).

[9] Our guess is that the model learned in Spanish can be adapted directly to Catalan.

4.2 Annotation Observations

From the data annotated (2,953 deverbal noun occurrences), we can highlight that implicit arguments in Spanish are more frequent than explicit arguments in nominal predicates. The average number of implicit arguments realized among the predicates analyzed, taking into account core and adjunct arguments, is almost two implicit arguments per instance (1.9). Therefore, the annotation of implicit arguments is crucial for the semantic treatment of deverbal nominalizations and implies a gain in role coverage of 317%[10]. Specifically, the core arguments arg0-agt/cau, arg1-pat/tem and arg2-ben/atr are those more frequently realized as implicit arguments.

Another relevant conclusion is that most implicit arguments are located nearby. From the total number of implicit arguments annotated, 60% are located within the sentence containing the nominal predicate, 32% are found within the previous context and 8% in the following context. Similar observations are drawn for English in (Gerber and Chai, 2012).

5 Conclusions

This project will give rise, on the one hand, to an enriched version of AnCora corpora with the annotation of the implicit arguments of deverbal nouns and, on the other hand, to the first available model of SRL dealing with implicit arguments in Spanish and Catalan.

IARG-AnCora will be the first corpus in these languages to be annotated with explicit and implicit arguments for deverbal noun predicates, with a high coverage available to the research community. This resource follows the same annotation scheme as NomBank and PropBank for argument structure, and as (Gerber and Chai, 2010; Gerber and Chai, 2012) for implicit arguments. In this way, we ensure the compatibility of the Spanish and Catalan resources with those that are also based on this annotation scheme. In fact, we aim to create interoperable semantic resources.

IARG-AnCora will be an important resource of semantic knowledge that could be used as a learning corpus for SRL nominal systems. It will also be a useful resource for linguistics studies on the argument structure of deverbal nominalizations or on coreference chains and the entities referring to NPs.

[10] This figure is extremelly higher than the reported for English (71%) in (Gerber and Chai, 2012) due to the lower degree of instantiation of explicit arguments.

References

Juan Aparicio, Mariona Taulé, and M.Antònia Martí. 2008. AnCora-Verb: A Lexical Resource for the Semantic Annotation of Corpora. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pages 797–802, Marrakech, Morocco, May. European Language Resources Association (ELRA).

Joseph L. Fleiss. 1981. *Statistical methods for rates and proportions*. John Wiley.

Matthew Gerber and Joyce Y. Chai. 2010. Beyond NomBank: a study of implicit arguments for nominal predicates. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1583–1592, Stroudsburg, PA, USA. Association for Computational Linguistics.

Matthew Gerber and Joyce Y Chai. 2012. Semantic Role Labeling of Implicit Arguments for Nominal Predicates. *Computational Linguistics*. To appear.

Matthew Gerber, Joyce Chai, and Adam Meyers. 2009. The role of implicit argumentation in nominal srl. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 146–154, Boulder, Colorado, June. Association for Computational Linguistics.

Jena D. Hwang, Archna Bhatia, Claire Bonial, Aous Mansouri, Ashwini Vaidya, Nianwen Xue, and Martha Palmer. 2010. Propbank annotation of multilingual light verb constructions. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 82–90, Uppsala, Sweden, July. Association for Computational Linguistics.

K. Kipper. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, Computer and Information Science Dept., University of Pennsylvania, PA.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of english: The penn treebank. *Computational Linguistics*, 19(2):313–330.

Adam Meyers, Ruth Reeves, and Catherine Macleod. 2004. NP-external arguments a study of argument sharing in English. In *Proceedings of the Workshop on Multiword Expressions: Integrating Processing (MWE*

'04), pages 96–103, Stroudsburg, PA, USA. Association for Computational Linguistics.

Adam Meyers. 2007. Annotation Guidelines for NomBank Noun Argument Structure for PropBank. Technical report, University of New York.

Lluis Márquez, Xavier Carreras, Kenneth C. Litkowski, and Suzanne Stevenson. 2008. Semantic role labeling: an introduction to the special issue. *Computational Linguistics*, 34(2):145–159.

Martha Palmer, Paul Kingsbury, and Daniel Gildea. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):76–105.

Martha Palmer, Daniel Gildea, and Nianwen Xue. 2010. *Semantic Role Labeling. Synthesis on Human Languages Technologies.* Morgan ang Claypool Piblishers.

Aina Peris and Mariona Taulé. 2011a. AnCora-Nom: A Spanish Lexicon of Deverbal Nominalizations. *Procesamiento del Lenguaje Natural.*, 46:11–19.

Aina Peris and Mariona Taulé. 2011b. Annotating the argument structure of deverbal nominalizations in Spanish. doi: 10.1007/s10579-011-9172-x. *Language Resources and Evaluation*.

Josef Ruppenhofer, Caroline Sporleder, Roser Morante, Collin Baker, and Martha Palmer. 2010. SemEval-2010 Task 10: Linking Events and Their Participants in Discourse. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 296–299, Uppsala, Sweden, July. Association for Computational Linguistics.

Mariona Taulé, M.Antónia Martí, and Marta Recasens. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, pages 96–101, Marrakech, Morocco, may. European Language Resources Association (ELRA).

Acknowledgments

This work is supported by the projects IARG-AnCora (FFI2011-13737-E), KNOW2 (TIN2009-14715-C04-04) and TextMess 2.0 (TIN2009-13391-C04-03/04) from the Spanish Ministry of Science and Innovation.