EusPropBank: Integrating Semantic Information in the Basque Dependency Treebank

Izaskun Aldezabal[1], María Jesús Aranzabe[1], Arantza Díaz de Ilarraza[2], Ainara Estarrona[2], and Larraitz Uria[3]

IXA NLP Group
[1] Basque Philology Department,
[2] Languages and Information Systems
University of the Basque Country
[3] IKER UMR 5478, University of Pau and Pays de l'Adour (UPPA), CNRS
{izaskun.aldezabal,maxux.aranzabe,a.diazdeilarraza,
ainara.estarrona,larraitz.uria}@ehu.es

Abstract. This paper deals with theoretical problems found in the work that is being carried out for annotating semantic roles in the Basque Dependency Treebank (BDT). We will present the resources used and the way the annotation is being done. Following the model proposed in the PropBank project, we will show the problems found in the annotation process and decisions we have taken. The representation of the semantic tag has been established and detailed guidelines for the annotation process have been defined, although it is a task that needs continuous updating. Besides, we have adapted AbarHitz, a tool used in the construction of the BDT, to this task.

Keywords: Theoretical Problems in Semantic Annotation, Representation of Semantic Roles, Lexical Resources.

1 Introduction

The construction of a corpus with annotation of semantic roles is an important resource for the development of advanced tools and applications such as machine translation, language learning and text summarization. We present here the work that is been carried out for annotating semantic roles in the BDT. Our previous work on semantics has mainly focused on word senses (including the development of the Basque WordNet and Basque Semcor (Agirre et al., 2006a), building verbal models from corpora, including selectional preferences (Agirre et al., 2003) and subcategorization frames (Aldezabal et al., 2003), as well as manually developing a database with syntactic/semantic subcategorization frames for a number of Basque verbs (Aldezabal, 2004).

Our interest follows the current trend, as shown by corpus tagging projects such us the Penn Treebank (Marcus, 1994), PropBank (Palmer et al., 2005) and PDT (Hajic et al., 2003), and the semantic lexicons that have been developed alongside them, suche as VerbNet (Kingsbury et al., 2002) and Vallex (Hajic et al., 2003). FrameNet (Baker et al., 1998) is also an example of the joint development of a semantic lexicon and a hand-tagged corpus.

A. Gelbukh (Ed.): CICLing 2010, LNCS 6008, pp. 60–73, 2010.

After a preliminary study, we chose to follow the PropBank/VerbNet model for a number of reasons:

- The PropBank project starts from a syntactically annotated corpus, just as we do.
- The organization of the lexicon is similar to our database of verbal models.
- Given the VerbNet lexicon and the annotations in PropBank, many implicit decisions on problematic issues, such as the distinctions between arguments and adjuncts have been settled and are therefore easy to replicate when we tag the Basque data.
- Having corpora in different languages annotated following the same model allows for cross-lingual studies and hopefully the enriching of Basque verbal models with the richer information currently available for English.

In fact, the PropBank model is being deployed in other languages, such as Chinese, Spanish, Catalan and Russian. Palmer and Xue (2003) and Xue (2008) describe the Chinese PropBank. Civit et al. (2005) describe a joint project to annotate comparable corpora in Spanish, Catalan and Basque.

The paper will be organised as follows: after a brief introduction, we will present the resources used in the semantic tagging. Section 3 explains the steps followed in the annotation, the automatic procedures defined to facilitate the task of manual annotation. In section 4, we describe the tool used for tagging (AbarHitz) while section 5 discusses theoretical problems and decisions we are facing. Finally, section 6 presents the conclusions and future work.

2 The Resources Used

In this section we will present the PropBank/VerbNet model, the model followed, and the resources we have for the annotation of semantic roles. We will explain them briefly, more details can be found in Aldezabal (2007) and Agirre et al. (2006b).

2.1 PropBank/VerbNet

PropBank is a corpus that is annotated with verbal propositions and their arguments. In the PropBank model two independent levels are distinguished: the level of arguments and adjuncts, and the level of semantic roles. The elements that are regarded as arguments are numbered from *Arg0* to *Arg5*, expressing semantic proximity with respect to the verb. The lowest numbers represent the main functions (subject, object, indirect object, etc.). The adjuncts are tagged as *ArgM*.

With regard to roles, PropBank uses two kinds: roles specific to each specific verb (e.g. buyer, thing bought, etc.), and general roles (e.g. agent, theme, etc.) linked to the VerbNet lexicon (Kipper et al., 2002).

VerbNet is an extensive lexicon where verbs are organized in classes following Levin's classification (1993). The lexicon provides an association between the syntactic and semantic properties of each of the described verbs.

Table 1 shows the PropBank roleset for the verb 'go.01' and the corresponding VerbNet roleset with Levin's class number (go-47.7 51.1-2).