Learning Argument/Adjunct Distinction for Basque Verbs

Abstract

This paper presents the experiments performed on lexical knowledge acquisition in the form of verbal subcategorization information. The system obtains the data from raw corpora after the application of a partial parser and statistical filters. We used three different statistical filters to acquire the subcategorization information: Mutual Information, Pearson's Chi-square and Fisher's Exact test.

Due to the characteristics of agglutinative languages like Basque, the usual classification of arguments in terms of their syntactic category (such as NP or PP) is not suitable. For that reason, the arguments will be classified in 48 different kinds of case markers, which makes the system fine grained if compared to equivalent systems that have been developed for other languages.

This work addresses the problem of distinguishing arguments from adjuncts, being this one of the most significant sources of noise in subcategorization frame acquisition.

1 Introduction

In the last years a considerable effort has been done on the acquisition of lexical information. As several authors point out, this information is useful for a wide range of applications. For example, Carroll and Minnen (1997) show how adding subcategorization information improves the performance of a parser.

Having this in mind our aim is to obtain a system that automatically discriminates between subcategorized elements of verbs (arguments) and non-subcategorized ones (adjuncts).

For evaluating our system we employed two different methods. The purpose was to find out which was the impact of each method on this particular task. We got significantly different results.

Basque is the subject of this study. A language that, in contrast to languages like English, has limited resources in the form of digital corpora, computational lexicons, grammars or annotated treebanks. Therefore, any effort like the one presented here, oriented to create lexical resources, has to be driven to do as much automatic work as possible, minimizing development costs.

The paper is divided into 5 sections. The second one is devoted to explain the theoretical motivations underlying the process. The third section is a description of the different stages of the system. The fourth section presents the results obtained under two different evaluation methods. It also discusses the possible causes of divergences between them. And finally, the fifth section is a review of previous work on automatic subcategorization acquisition. Finally, we present the main conclusions.

2 The argument/adjunct distinction

Talking about Subcategorization Frames (SCF), means talking about arguments. Many existing systems acquire directly a set of possible SCFs without any previous filtering of adjuncts. This is a substantial source of noise. Therefore, in order to avoid this problem, our approach is based on the argument/adjunct distinction.

The argument/adjunct distinction is probably one of the most unclear issues in linguistics. Many authors consider the distinction to be a continuum where some elements belonging to the extremes of this continuum can be easily classified as arguments or adjuncts, while those belonging to the central part of the continuum can be easily misclassified. From the different diagnostics proposed some are quite consistent among the different authors (Grishman et al. 1994).

1) The Obligatoriness condition. When a verb demands obligatorily the appearance of an element, this element will be an argument.
2) Frequency. Arguments of a verb occur more frequently with that verb than with the other verbs.

These two diagnostics can be captured statistically by the application of association measures like Mutual Information.

1) … EEBBetako lehendakariak UEko 15 herrialdeetako merkataritza ministroekin bazkaldu behar zuen negoziazioen bilgunean …

2) … the president of the USA had to eat with the ministers of Commerce of 15 countries of the UE in the negotiation center …

3.1) [EEBB-etako lehendakari-a-k] [UE-ko 15 herrialde-etako merkataritza ministro-ekin] →
[USA-of president-the-erg.] [UE-of 15 countries-of Commerce ministers-with]
3.2) NP-ergative(president, sing) PP(with)-sociative(minister, plural)

→ *[bazkaldu behar zuen]* *[negoziazio-en bilgune-an]*
[to eat had] [negotiation-of center-in]
verb(eat) PP(in)-inessive(center, sing)

Figure 1. Example of the output of the shallow parsing phase: 1) Input (in Basque), 2) English translation, 3.1) Verb and dependents (phrases), and 3.2) Verb and dependents (case+head)

Before going into a discussion about the different measures that could be applied, we will present step by step the whole process we pursued for achieving the argument/adjunct distinction.

3 The acquisition process

Our starting point was a raw corpus of 1,337,445 words, where there were instances of 1,412 verbs. From them, we selected 640 as statistically relevant because they appear in more than 10 sentences.
As we said earlier, our goal was to distinguish arguments from adjuncts. When starting from raw corpus, like in this case, it is necessary to get instances of verbs together with their dependents (arguments and adjuncts). We got this information applying a partial parser (section 3.1) to the corpus. Once we had the dependents, statistical measures helped us deciding which were arguments and which were adjuncts (section 3.2).

3.1 The parsing phase

Aiming to obtain the data against which statistical filters will be applied, we analyzed the corpus using several available linguistic resources:

- First, we performed the morphological analysis of the corpus, based on two-level morphology (Koskenniemi 1983; Alegria et al. 1996) and disambiguation using the Constraint Grammar formalism (Karlsson et al. 1995, Aduriz et al. 1997).
- Second, a shallow parser was applied (Aldezabal et al. 2000), which recognizes basic syntactic units including noun phrases, prepositional phrases and several types of subordinate sentences.
- The third step consisted in linking each verb and its dependents. Basque lacks a robust parser for performing full parsing as in (Briscoe and Carroll 1997, Kawahara et al. 2001), and therefore we used a finite state grammar (Aldezabal et al. 2001) to link the dependents (both arguments and adjuncts) with the verb. This grammar was developed using the Xerox Finite State Tool (Karttunen et al. 1997). Figure 1 shows the result of the parsing phase. In this case, both sociative and inessive cases (PPs) are adjuncts.
  The linking of dependents to a verb is not trivial considering that Basque is a free word order language, and any element appearing between two verbs could be, in principle, dependent on any of them. Many problems must be taken into account, such as ambiguity and determination of clause boundaries, among others. We evaluated the accuracy up to this point and we got a precision of 87% and a recall of 66%. So the input data to the next phase was relatively noisy.

3.2 The argument selection phase

In the data resulting from the shallow parsing phase we counted up to 65 different cases (types of arguments, including postpositions and different types of suffixes). These are divided in two main groups:

- 43 correspond to postpositions. Some of them can be directly mapped to English prepositions, but in many cases several Basque postpositions correspond to just one English preposition (see Table 1). This set also contains postpositions that map to categories other than English prepositions, such as adverbs.

Table 1. Correspondence between English prepositions and Basque postpositions.

English	Basque
To	Dative Alative Final ablative
Like	-en gisa gisa bezala legez

- 22 types of sentential complements (For instance, English *that* complementizer corresponds to several subordination suffixes: *-la, -n, -na, -nik*).

This shows to which extent the range of arguments is fine-grained, in contrast to other works where the range is at the categorial level, such as NP or PP (Brent 1993, Manning 1993, Merlo 2001).

Due to the complexity carried by having such a high number of cases, we decided to gather postpositions that are equivalent or almost equivalent (for example, English *between* and *among*). Even if there are some semantic differences between them they do not seem to be relevant at the syntactic level. Some linguists were in charge of completing this grouping task. Even considering the risk of making mistakes when grouping the cases, we concluded that the loss of accuracy due to having too sparse data (consequence of having many cases) would be worse than the noise introduced by any mistake in the grouping. The resulting set contained 48 cases. The complexity got reduced but it is still considerable.

Another important aspect we want to mention is that our goal is to get information about the elements that could be either arguments or adjuncts. In Basque there are three structural cases which are special because, when they appear, they are always arguments. They correspond to the subject, direct object and indirect object functions. These cases are not very conflictive when parsing. The reason is that in Basque the auxiliary bears information about their appearance in the sentence. So they are easily recognized and linked to the corresponding verb. That is the reason for not including them in this work. Probably Precision and Recall would improve if we decided to include them, because they are the most frequent cases in the corpus, and statistics perform well over frequent data.

Most of the works on automatic acquisition of subcategorization information (Carroll & Briscoe 1997, Sarkar & Zeman 2000, Korhonen 2001) apply statistical methods (hypothesis testing). Basically the idea is the following: they get "possible subcategorization frames" from automatically parsed data (either completely or partially parsed) or from a syntactically annotated corpus. Afterwards a statistical method is employed to decide whether those "possible frames" are or not real subcategorization frames. These statistical methods can be problematic mostly because they perform badly on sparse data. In order to avoid as much as possible data sparseness, we decided to design a system that learns which are the arguments of a given verb instead of learning whole frames. Frames are combinations of arguments, and considering that our system deals with 48 cases, the number of combinations was high, resulting in sparse data. So we decided to work at the level of the argument/adjunct distinction. Working on this distinction is also very useful to avoid noise in the subcategorization frame, because adjuncts are synonyms of noise in the acquisition of subcategorization frames. A system that tries to get subcategorization frames without previously making the argument/adjunct distinction will suffer of having highly sparse and noisy data.

To accomplish the argument/adjunct distinction we applied three statistical measures: Mutual Information (MI), Pearson's Chi-square and Fisher's Exact Test (for more information on these measures, see Manning & Hinrich 1999). All these tests are broadly used to discover associations between words, but they show different behaviour depending on the nature of the data (Dunning 93). MI seems to overestimate on small number of occurrences. Pearson's Chi-square is based on the assumption of normality. That is to say, it is only applicable when the data used is normally distributed. And, finally, there is agreement on the good behaviour of Fisher´s Exact Test. We did not want to make any a priori decision on the measure employed. On the contrary, we aimed to check which test was behaving better on our data.

4 Evaluation

We found in the literature two main approaches to evaluate a system like the one proposed in this paper (Briscoe & Carroll 1997, Sarkar & Zeman 2000, Korhonen 2001):

- Comparing the obtained information with a gold standard.
- Calculating the coverage of the obtained information on a corpus.

Under the former approach a further distinction emerges: some systems use a dictionary as a gold standard. Some others perform manual evaluation by extracting the subcategorization frames appearing in a corpus and comparing them with the set of subcategorization frames obtained automatically. We decided to do it both ways, that is to say, using a gold standard and calculating the coverage over a corpus. The intention was to determine, all things being equal, the impact of doing it one way or the other.

4.1 Evaluation 1: comparison of the results with a gold standard

From the 640 analyzed verbs, we selected 10 for evaluation. For each of these verbs we extracted from the corpus the list of all their dependents. We provided 4 human annotators with this list and they marked each dependent as either argument or adjunct. The list obtained from merging[1] the 4 lists in one is taken to be our gold standard. Notice that when the annotators decided whether a possible argument was really an argument or not, no context was involved. In other words, they were deciding over bare pairs of verbs and cases. Therefore different senses of the verb were considered, even those not appearing in the corpus, because there was no context to disambiguate the specific meaning of the verb. Table 1 shows the results obtained in terms of Precision and Recall.

Table 1. Results of Evaluation 1 (Context independent)

	Precision	Recall	F-score
MI	62%	50%	55%
Pearson	24%	56%	34%
Fisher	64%	44%	52%

[1] Merging was possible once the annotators agreed on the marking of each element.

4.2 Evaluation 2: Calculation of the coverage on a corpus

The initial corpus was divided in two parts, one for training the system and another one for evaluating it. From the fraction reserved for evaluation we extracted 200 sentences corresponding to the same 10 verbs used in the "gold standard" based evaluation. In this case, the task carried out by the annotators consisted in extracting, for each of the 200 sentences, the elements (arguments/adjuncts) linked to the corresponding verb. Each element was marked as argument or adjunct with respect to the corresponding verb. Note that in this case the annotation takes place inside the context of the sentence, in other words, the verb shows precise semantics.

We performed a simple evaluation on the sentences, calculating precision and recall over each argument marked by the annotators. For example, if a verb appeared in a sentence with two arguments and the statistical filters were recognizing them as arguments, both precision and recall would be 100%. If, on the contrary, only one was found, then precision would be 100%, and recall 50%.

Table 2. Results of Evaluation 2 (inside context)

	Precision	Recall	F-score
MI	93%	97%	95%
Pearson	93%	61%	74%
Fisher	93%	93%	93%

4.3 Discussion

It is obvious that the results attained in the first evaluation are worse than those in the second one. The origin of this difference comes, on one hand, from semantics and, on the other hand, from the nature of statistics.

- Semantic source. The former experiment was not contextualized, while the latter used a context. Our experience showed us that broader semantics (non-contextualized) leads to a situation where the number of arguments increases with respect to narrower (contextualized) semantics. This happens because in many cases different senses of the same verb require different arguments. Therefore, the task becomes more difficult. In addition to this, and also tied to semantics, the corpus used for learning and evaluating were different parts of the same corpus, with similar, if not equal, semantics.

- Statistical origin. The disagreement in the results comes from the nature of the statistics themselves. Any statistical measure performs better on the most frequent cases than on the less frequent ones. In the first experiment, all possible arguments are evaluated, including the less frequent ones, whereas in the second experiment, only the possible arguments found in the piece of corpus used were evaluated. In most of the cases, the possible arguments found were the most frequent ones.

5 Related work

Concerning the acquisition of verb subcategorization information, there are proposals ranging from manual examination of corpora (Grishman et al. 1994) to fully automatic approaches.

Table 3, partially borrowed from Korhonen (2001), summarizes several systems on subcategorization frame acquisition.

Manning (1993) presents the acquisition of subcategorization frames from unlabelled text corpora. He uses a stochastic tagger and a finite state parser to obtain instances of verbs with their adjacent elements (either arguments or adjuncts), and then a statistical filtering phase produces subcategorization frames (from a set of previously defined 19 frames) for each verb.

Briscoe and Carroll (1997) describe a grammar based experiment for the extraction of subcategorization frames with their associated relative frequencies, obtaining 76.6% precision and 43.4% recall. Regarding evaluation, they use ANLT and COMLEX Syntax dictionaries as gold standard. They also performed evaluation of coverage over a corpus. For our work, we could not make use of any previous information on subcategorization, because there is nothing like a subcategorization dictionary for Basque.

Sarkar and Zeman (2000) report results on the automatic acquisition of subcategorization frames for verbs in Czech, a free word order language. The input to the system is a set of manually annotated sentences, where each verb is linked with its dependents (without distinguishing arguments and adjuncts). The task consists in iteratively eliminating elements from the possible frames with the aim of removing adjuncts. For evaluation, they give an estimate of how many of the obtained frames appear in a set of 500 sentences where dependents were annotated manually, showing an improvement from a baseline of 57% (all elements are adjuncts) to 88%.

Comparing this approach to our work, Sarkar and Zeman's data does not come from raw corpus, thus they do not deal with the problem of noise coming from the parsing phase. Their main limitation comes by relying on a treebank, which is an expensive resource.

Kawahara et al. (2001) use a full syntactic parser to obtain a case frame dictionary for Japanese, where arguments are distinguished by their syntactic case, including their headword (selectional restrictions). The resulting case frame components are selected by a frequency threshold.

Maragoudakis et al. (2001) apply a morphological analyzer and phrase chunking module to acquire subcategorization frames for Modern Greek. In contrast to this work, they use different machine learning techniques. They claim that Bayesian Belief Networks are the best learning technique.

Merlo (2001) presents supervised learning experiments for automatic distinction of arguments and adjuncts, applied to the case of prepositional phrases attached to a verb. She uses decision trees tested on a set of 400 verb instances with a single PP, reaching an accuracy of 86.5% over a baseline of 74%.

Note that both Manning and Merlo's systems learn from contexts with just one PP (maximum) per verb (finite state filter). Our system learns from contexts with up to 5 PPs. Furthermore, we distinguish 48 different kinds of cases, hence the combination number increases.

Regarding the parsing phase, the systems presented so far are heterogeneous. While Manning, Merlo and Maragoudakis et al. use very simple parsing techniques, Briscoe & Carroll and Kawahara et al. use sophisticated parsers. Our system can be placed between these two approaches. The result of the shallow parsing is not simple in that it relies on a robust morphological analysis and disambiguation. Remember that Basque is agglutinative language with strong morphology and, therefore, this stage is particularly relevant. Moreover, the finite state filter used for parsing is very sophisticated (Karttunen et al. 1997).

6 Conclusion

This work describes an initial effort to obtain subcategorization information for Basque. To successfully perform this task we had to go deeper than mere syntactic categories (NP, PP, …) enriching the set of possible arguments to 48 different classes. This leads to quite sparse data. Together with sparseness, another problem common to every subcategorization acquisiton system is that of noise, coming from adjuncts and incorrectly parsed elements. For that reason, we defined subcategorization acquisition in terms of distinguishing between arguments and adjuncts.

We carried out two different kinds of evaluation, which allowed us to verify the relevance of semantics in this kind of task.

For the future, we plan to incorporate the information resulting from this work in our parsing system. We hope that this will lead to get better results in parsing. Consequently, we would get better subcategorization information (bootstrapping). We also plan to improve the results by using semantic information as Korhonen (2001) proposes. And finally we also intend to use a bigger quantity of data.

Acknowledgements

This work was partly supported by a grant provided by the Basque Government.

Table 3. Summary of several systems on subcategorization frame acquisition

Method	Number of frames	Number of verbs	Linguistic resources	F-Score (evaluation based on a gold standard)	Coverage on a corpus
Manning (1993)	19	200	POS tagger + simple finite state parser	58	
Briscoe & Carroll (1997)	161	14	Full parser	55	
Sarkar & Zeman (2000)	137	914	Annotated treebank	-	88
Kawahara et al. (2001)	-	23,497	Full parser		82 accuracy
Maragoudakis et al. (2001)	-	47	Simple phrase chunker	77	
This paper	-	640	Morph. Analyzer + Phrase Chunker + Finite State Parser	55	95

References

Aduriz I., Arriola J.M., Artola X., Díaz de Ilarraza A., Gojenola K. and Maritxalar M. (1997) *Morphosyntactic disambiguation for Basque based on the Constraint Grammar Formalism.* Conference on Recent Advances in Natural Language Processing (RANLP).

Alegria I., Artola X., Sarasola K., Urkia M. (1996) *Automatic morphological analysis of Basque.* Literary and Linguistic Computing. 11 (4), Oxford University.

Aldezabal I., Gojenola K., Sarasola K. (2000) *A Bootstrapping Approach to Parser Development.* International Workshop on Parsing Technologies (IWPT), Trento.

Aldezabal I., Aranzabe M., Atutxa A., Gojenola K., Oronoz M., Sarasola K. (2001) *Application of finite-state transducers to the acquisition of verb subcategorization information.* Finite State Methods in Natural Language Processing, ESSLLI Workshop, Helsinki.

Brent, M. R. (1993) *From Grammar to Lexicon: Unsupervised Learning of Lexical Syntax.* Computational Linguistics, 19:243-262.

Briscoe, T. and Carroll J. (1997) *Automatic Extraction of Subcategorization from Corpora.* ANLP-97:356-363.

Carroll J., Minnen G. and Briscoe T. (1998) *Can Subcategorization Probabilities Help a Statistical Parser?* Proceedings of the 6th ACL/SIGDAT Workshop on Very Large Corpora, Montreal.

Karlsson F., Voutilainen A., Heikkila J., Anttila A. (1995) *Constraint Grammar: A*

Language-independent System for Parsing Unrestricted Text. Mouton de Gruyter.

Karttunen L., Chanod J-P., Grefenstette G., Schiller A. (1997) *Regular Expressions For Language Engineering*. Natural Language Engineering.

Kawahara D., Kaji N. and Kurohashi S. (2000) *Japanese Case Structure Analysis by Unsupervised Construction of a Case Frame Dictionary*. COLING-2000, Saarbrucken.

Korhonen A. (2001) *Subcategorization acquisition*. PhD Thesis, University of Cambridge.

Koskenniemi K. (1983) *Two-level Morphology: A general Computational Model for Word-Form Recognition and Production*. PhD thesis, University of Helsinki.

Kuhn J., Eckle-Kohler J., Rohrer. C. (1998) *Lexicon Acquisition with and for Symbolic NLP-Systems -- a Bootstrapping Approach*. First International Conference on Language Resources and Evaluation (LREC98), Granada.

Grishman R., Macleod C., Meyers A. (1994) *Comlex Syntax: Building a Computational Lexicon*. COLING-94.
Karttunen L., Chanod J-P., Grefenstette G., Schiller A. (1997) *Regular Expressions For Language Engineering*. Natural Language Engineering.
Manning, C. D. (1993) *Automatic Acquisition of a Large Subcategorization Dictionary from Corpora*. Proceedings of the 31th ACL.

Maragoudakis M., Kermanidis K., Fakotakis N. and Kokkinakis G. (2001) *Learning Automatic Acquisition of Subcategorization Frames using Bayesian Inference and Support Vector Machines*. The 2001 IEEE International Conference on Data Mining, IMDC'01, San José.

Merlo P. and Leybold M. (2001) *Automatic Distinction of Arguments and Modifiers: the Case of Prepositional Phrases*. EACL-2001, Toulousse.

Sarkar, A. and Zeman D. (2000) *Automatic Extraction of Subcategorization Frames for Czech*. COLING-2000, Saarbrucken.